Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique

François Morlane-Hondère

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse II Le Mirail (UT2 Le Mirail)

Discipline ou spécialité :

Sciences du langage

Présentée et soutenue par :
François Morlane-Hondère

le : mercredi 10 juillet 2013

Titre :

Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique

---

Ecole doctorale :
Comportement, Langage, Education, Socialisation, COgnition (CLESCO)

Unité de recherche :
CLLE-ERSS

Directeur(s) de Thèse :

Cécile Fabre - Professeur, Université de Toulouse II/CLLE-ERSS

Rapporteurs :

Béatrice Daille - Professeur, Université de Nantes/LINA
Alain Polguère - Professeur, Université de Lorraine/ATILF

Membre(s) du jury :

Nabil Hathout - Directeur de recherche, CNRS/CLLE-ERSS
Pierre-André Buvet - Maître de conférences HDR, Université Paris 13/LDI

Résumé

Dans cette thèse, nous abordons du point de vue linguistique la question de l'évaluation des bases lexicales extraites par analyse distributionnelle automatique (ADA). Les méthodes d'évaluation de ces ressources qui sont actuellement mises en oeuvre (comparaison à des lexiques de référence, évaluation par la tâche, test du TOEFL...) relèvent en effet d'une approche quantitative des données qui ne laisse que peu de place à l'interprétation des rapprochements générés. De ce fait, les conditions qui font que certains couples de mots sont extraits alors que d'autres ne le sont pas restent mal connues. Notre travail vise une meilleure compréhension des fonctionnements en corpus qui régissent les rapprochements distributionnels. Pour cela, nous avons dans un premier temps adopté une approche quantitative qui a consisté à comparer plusieurs ressources distributionnelles calculées sur des corpus différents à des lexiques de références (le Dictionnaire électronique des synonymes du CRISCO et le réseau lexical JeuxDeMots). Cette étape nous a permis, premièrement, d'avoir une estimation globale du contenu de nos ressources, et, deuxièmement, de sélectionner des échantillons de couples de mots à étudier d'un point de vue qualitatif.

Cette deuxième étape constitue le coeur de la thèse. Nous avons choisi de nous focaliser sur les relations lexico-sémantiques que sont la synonymie, l'antonymie, l'hyperonymie et la méronymie, que nous abordons en mettant en place quatre protocoles différents. En nous appuyant sur les relations contenues dans les lexiques de référence, nous avons comparé les propriétés distributionnelles des couples de synonymes/antonymes/hyperonymes/méronymes qui ont été extraits par l'ADA avec celles des couples qui ne l'ont pas été. Nous mettons ainsi au jour plusieurs phénomènes qui favorisent ou bloquent la substituabilité des couples de mots (donc leur extraction par l'ADA). Ces phénomènes sont considérés au regard de paramètres comme la nature du corpus qui a permis de générer les bases distributionnelles étudiées (corpus encyclopédique, journalistique ou littéraire) ou les limites des lexiques de référence.

Ainsi, en même temps qu'il questionne les méthodes d'évaluation des bases distributionnelles actuellement employées, ce travail de thèse illustre l'intérêt qu'il y a à considérer ces ressources comme des objets d'études linguistiques à part entière. Les bases distributionnelles sont en effet le résultat d'une mise en oeuvre à grande échelle du principe de substituabilité, ce qui en fait un matériau de choix pour la description des relations lexico-sémantiques.

Remerciements

De la même façon que le sens d'un mot est influencé par son contexte, une thèse se construit dans un environnement aussi bien universitaire qu'extra-universitaire qu'il serait inconcevable de ne pas évoquer ici. Ce travail est dédié à toutes les personnes dont les qualités diverses ont contribué à donner du sens à ces quatre années de thèse et à faire de moi un doctorant [+ EUPHORIQUE] et [+ ÉPANOUI].

Mes remerciements s'adressent tout d'abord à ma directrice, Cécile Fabre, avec qui j'ai la chance de travailler depuis ma première année de master. Je lui dois énormément car c'est elle qui m'a incité à me lancer dans cette aventure qu'est la thèse et qui, par sa disponibilité et sa pédagogie, a fait en sorte que cette aventure se déroule le plus sereinement possible. Ses nombreuses relectures et la pertinence de ses conseils ont grandement contribué à faire de ce mémoire de thèse un document dont je suis fier. Je la remercie ici de la confiance qu'elle m'a accordée, pour son investissement dans mon travail et, plus généralement, pour avoir su me donner le goût de la recherche.

Je remercie chaleureusement les membres de mon jury, à savoir Béatrice Daille et Alain Polguère pour avoir accepté d'être les rapporteurs de ce travail, ainsi que Nabil Hathout et Pierre-André Buvet, qui ont bien voulu en être les examinateurs.

Cette thèse a été réalisée au sein du laboratoire CLLE-ERSS, dont je tiens à remercier les membres pour leur disponibilité et leur bienveillance. J'adresse une pensée particulière à Myriam Bras, Anne Condamines, Cécile Fabre, Marie-Paule Péry-Woodley, Josette Rebeyrolle, Nathalie Rossi, Ludovic Tanguy et Jean-Michel Tarrier, dont j'ai suivi les cours à l'époque où j'étais *de l'autre côté* et qui m'ont donné envie d'aller plus loin dans l'étude de la linguistique, et plus particulièrement dans son versant TAL. Je remercie également Franck Sajous pour m'avoir fourni en voisins distributionnels ainsi que Mai Ho-Dac, avec qui j'ai eu le plaisir de partager un TD ces deux dernières années.

Je ne remercierai également jamais assez les doctorants et anciens doctorants que j'ai eu l'occasion de côtoyer au sixième étage durant ces quatre années. Ils sont tous devenus des amis et travailler à leurs côtés a été un réel plaisir. Je remercie donc :

– ceux qui sont partis vers de nouveaux horizons (Aurélie Picton, Buddy Dirat, Christelle Pêcher) ou qui sont juste descendus d'un étage (Marianne Vergez-Couret et Clémentine Adam, mes anciennes collègues de bureau, que je remercie de m'avoir si bien acueilli à mon arrivée au laboratoire) ;

– ceux avec qui j'ai eu l'occasion de partager une place dans le bureau C617, à savoir Marianne Vergez-Couret et Clémentine Adam (bis), Simon Leva, Jean-Philippe Fauconnier, Marin "Ce n'est pas acceptable !" Popan, Assaf Urieli, Lama Allan et Matteo Pascoli ;
– Fanny Lalleman, Nikola Tulechki et Marjorie Raufast (doctorante potentielle), que j'ai eu l'occasion de côtoyer depuis la licence et qui ont eux aussi adopté la voie du TAL ;
– l'ensemble des membres – passés, actuels et futurs – du Cercle d'Étudiants pour l'Étude du Langage (CEPEL), en particulier Charlotte Alazard, qui a été une excellente co-responsable ;
– les camarades de rédaction des week-ends, Marie-France Roquelaure et Céline Launay (qui a bien voulu nous accueillir dans son bureau le jour où un mystérieux inconnu a fait sauter les plombs dans les bureaux des doctorants) ;
– Céçaïle Viollain et Marine Lasserre, représentantes des doctorants, pour tous les efforts qu'elles fournissent pour égayer le quotidien du sixième étage (la chasse aux œufs de Pâques restera dans les mémoires !) ;
– tous les autres doctorants, dont beaucoup sont sur le point de soutenir : Caroline Atallah, Nathalie Dehaut, Aurélie Guerrero, Luce Lefeuvre, Stéphanie Lopez, Florian Savreux et Caitlin Smith. J'envoie également un "monstre" remerciement à Laure-Anne Johnsen pour avoir apporté le temps de quelques mois sa bonne humeur à Toulouse ;
– Guillaume Carbou, pour énormément de choses : les assauts libres des cours de savate, les débats du midi à la piste *plat du monde*, les cafés de la machine, les soupes auto-chauffantes à l'oxyde de calcium, les aveugles qui ont un frère – alors que celui-ci n'en a pas ! – et, plus généralement, pour les milliers de crises de rire depuis qu'on s'est retrouvés en salle informatique à annoter des structures énumératives.

Dans le cadre extra-universitaire, j'aimerais remercier ceux qui m'ont permis de m'aérer le corps et l'esprit, à savoir mes compagnons de concert Julien "Flower power" Clariond, Julie Diaz, Fabien Hauret-Clos, Caroline Muniesa, Caroline Vidal, et John Vidal (avec qui j'ai assisté à mon premier cours de linguistique à l'université de Pau), ainsi que mes camarades des randos roller du vendredi soir Ludovic Campion, Sarah Médous, Amaury Thiebault et David Wagner.

Je remercie bien sûr l'ensemble de ma famille et particulièrement mes parents, pour le soutien sans faille qu'ils m'ont apporté depuis le début de mon cursus universitaire. Je suis également reconnaissant envers mon cousin Pierre Sarrailh, que j'ai suivi sur le chemin de la thèse, pour tous ses précieux conseils dispensés autour d'une pizza.

Enfin, je remercie Camille, qui a supporté sans broncher le rythme de vie infernal imposé par la rédaction ces quatre derniers mois. Sans ses attentions et ses encouragements constants, cette fin de thèse aurait certainement été autrement plus éprouvante. Je lui témoigne ici ma reconnaissance et mon amour.

Table des matières

Liste des abréviations

AD : analyse distributionnelle.

ADA : analyse distributionnelle automatique.

BNC : British National Corpus.

DES : Dictionnaire Électronique des Synonymes.

IM : information mutuelle.

JDM : JeuxDeMots.

TAL : traitement automatique des langues.

TLF : Trésor de la Langue Française.

VDF : voisins de Frantext.

VDLM : voisins de Le Monde.

VDW : voisins de Wikipédia.

Introduction

L'analyse distributionnelle (AD) s'appuie sur un principe simple, celui d'une corrélation entre les contextes dans lesquels les mots apparaissent – leur *distribution* – et leur contenu sémantique. Le corollaire de ce principe est que la distance sémantique entre deux mots est proportionnelle à la quantité de contextes qu'ils partagent. De ce fait, l'observation des contextes dans lesquels apparaissent les mots d'un corpus permet de mettre au jour des classes sémantico-distributionnelles.

L'automatisation de ce principe énoncé par Z. Harris a été motivée par le besoin de disposer à moindre coût de ressources sémantiques *à la WordNet* destinées à être intégrées à des systèmes de traitement automatique des langues (TAL). En effet, l'approche consistant à mobiliser des experts pour construire un réseau sémantique s'avère extrêmement coûteuse en temps et en moyens. L'analyse distributionnelle automatique (ADA), de par son potentiel à extraire des relations de sens à partir de grands volumes de textes tout en ne nécessitant que peu d'intervention humaine, est alors apparue comme une alternative séduisante. Dans cette approche, le réseau des relations distributionnelles émerge automatiquement de l'analyse des textes : par exemple, dans un corpus composé d'articles de Wikipédia, *chêne* apparaît de façon récurrente dans des contextes comme *écorce de* $\sim$, *espèce de* $\sim$ ou en position objet du verbe *planter*, ce qui est également le cas de *pin* ou de *arbre*, lequel est rapproché de *forêt* et *bois* par les contextes *exploitation de* $\sim$, *vivre dans* $\sim$, *esprit de* $\sim$, etc.

Toutefois, à l'heure actuelle, les ressources que produit l'ADA sont encore loin de pouvoir constituer des thesaurus utilisables en l'état. La raison en est double :

- la relation de proximité distributionnelle qu'entretiennent les couples produits par l'ADA n'est pas caractérisée du point de vue sémantique. Or, cette relation est ambigüe : deux mots rapprochés par l'ADA peuvent entretenir un vaste ensemble de relations de sens, parmi lesquelles des relations dites *classiques* comme la synonymie, l'antonymie, l'hyperonymie, etc., des relations *non classiques* (relations syntagma-

tiques, thématiques) ou encore des relations pour lesquelles la proximité distributionnelle ne correspond pas à une relation sémantique reconnaissable intuitivement ;
- les rapprochements que produit cette méthode sont pléthoriques : les ressources que nous utilisons ont été produites à partir d'une méthode qui extrait jusqu'à 5,5 millions de couples à partir d'un corpus d'environ 200 millions de mots. Un mot est en moyenne rapproché de 147 autres (ce chiffre peut aller jusqu'à plusieurs milliers dans certains cas).

Ainsi, si les ressources distributionnelles possèdent l'avantage de la quantité – en termes de données traitées et produites –, la question de la qualité reste en suspens. Les méthodes qui sont utilisées actuellement pour évaluer le contenu des ressources distributionnelles posent également problème. L'une d'entre elles consiste à comparer les données produites par l'ADA à des ressources de référence, le plus souvent à des dictionnaires ou des réseaux lexicaux contenant des couples de synonymes, antonymes, etc. La qualité de la base distributionnelle évaluée correspond ainsi à la proportion de couples de la ressource de référence qu'elle contient. Un autre mode d'évaluation consiste à intégrer une base distributionnelle à un système de TAL et à jauger la qualité de la ressource à l'aune des performances de ce système. Une troisième approche largement répandue consiste à utiliser les ressources distributionnelles pour accomplir des tâches lexicales comme la reconnaissance de synonymes (typiquement avec les tests du TOEFL). Ces trois méthodes ne répondent cependant que de façon superficielle à la question du contenu des bases distributionnelles, dans le sens où elles ne permettent pas d'expliquer les fonctionnements linguistiques qui sont à l'œuvre dans les ressources évaluées.

L'approche que nous adoptons dans ce travail de thèse se distingue de celles que nous avons évoquées plus haut par le fait qu'elle repose sur une conception des bases distributionnelles comme des objets linguistiques. Ce point de vue, loin d'être incompatible avec les modes d'évaluation traditionnels, s'en veut complémentaire : la comparaison d'une base distributionnelle avec un lexique de référence n'est alors plus considérée comme une fin en soi mais comme la première étape d'une analyse qualitative des données. Nous nous situons en effet dans une approche à l'interface entre la lexicologie et la TAL qui consiste, dans un premier temps, à se doter d'un ensemble de couples porteurs d'une relation sémantique donnée – issus de lexiques ou extraits avec des patrons – que nous croisons avec des ressources distributionnelles. Nous nous appuyons ensuite sur l'étude comparée des contextes d'apparition des mots qui composent ces couples pour expliquer les raisons qui font que certains ont été extraits par l'ADA alors que d'autres non. L'originalité de cette

démarche réside d'une part dans le fait que nous apportons de la linguistique là où elle est réduite à la portion congrue et, réciproquement, dans le fait que nous montrons l'intérêt d'un outil trop peu usité sur l'établi de la linguistique outillée. Les apports que cette approche peut offrir se situent donc à deux niveaux :

- le fait d'améliorer la compréhension des fonctionnements qui influent sur le type de couples qui sont captés dans les bases distributionnelles participera à l'identification des limites et du potentiel de ces ressources. Cela permettra d'en faire une utilisation plus éclairée ;
- en montrant comment certains principes traditionnellement admis en lexicologie peuvent être questionnés à la lumière des données contenues dans les bases distributionnelles, nous faisons la démonstration du potentiel de ces ressources pour la description des relations lexicales. Ces dernières se définissent en effet traditionnellement par rapport au principe de substituabilité, dont l'ADA constitue une mise en œuvre à grande échelle. Elles peuvent donc permettre au lexicologue d'éprouver ce principe de substituabilité en s'appuyant sur des données calculées à partir de corpus de différentes natures.

Ce document est organisé en deux parties. La partie I, intitulée simplement *L'analyse distributionnelle automatique*, englobe les trois premiers chapitres. Elle se veut à la fois un historique et un état de l'art de l'ADA. Dans le chapitre 1, nous revenons à l'origine de l'AD en évoquant ses principes fondateurs à travers les travaux de Harris. Nous évoquons également les difficultés qui ont été rencontrées lors des premières tentatives d'automatisation de la méthode. Dans le chapitre 2, nous décrivons successivement les différents paramètres qui entrent en jeu lors de la mise en œuvre de l'ADA. En nous appuyant sur la littérature, nous décrivons l'influence que peut avoir chacun de ces paramètres sur la composition des ressources distributionnelles générées. Dans le chapitre 3 nous faisons un panorama des modèles distributionnels existants avant de décrire le modèle Syntex-Upéry, qui a permis de générer les ressources distributionnelles avec lesquelles nous travaillons dans les expériences décrites dans la partie suivante.

La partie II constitue la partie expérimentale de notre travail. Elle s'ouvre sur le chapitre 4, dans lequel nous évoquons les types de relations que l'ADA permet de mettre au jour et les limites des méthodes les plus couramment utilisées pour évaluer les ressources distributionnelles. Nous employons toutefois l'une de ces méthodes – la comparaison à des dictionnaires constitués manuellement – afin de mettre en évidence des phénomènes de décalages distributionnels à grande échelle que nous observons au niveau des contextes d'apparition des mots dans les chapitres suivants. Les principes méthodologiques qui guident notre approche sont décrits dans un préambule qui s'in-

sère entre les chapitres 4 et 5. Et bien que cette démarche de questionner la réciprocité entre relation sémantique et relation distributionnelle guide l'ensemble des expériences que nous menons par la suite, nous avons choisi de faire varier la méthodologie d'observation en fonction des relations abordées. Les relations que nous étudions sont les suivantes :

- la synonymie (chapitre 5) : nous cherchons à montrer que le fait de filtrer successivement un dictionnaire de synonymes avec plusieurs bases distributionnelles permet de mettre en lumière les synonymes les plus pertinents pour un type de texte donné ;
- l'antonymie (chapitre 6) : nous abordons cette relation en adoptant un point de vue inspiré des approches psycholinguistiques. Nous faisons l'hypothèse que le fait de croiser des couples de mots extraits par des patrons antonymiques et une base distributionnelle peut mettre au jour des couples antonymiques dont le double fonctionnement à la fois syntagmatique et paradigmatique serait gage d'un degré particulièrement élevé d'antonymie ;
- l'hyperonymie (chapitre 7) : nous cherchons à montrer que l'observation des liens de similarité distributionnelle entre un hyperonyme et ses hyponymes met en lumière des différences dans la catégorisation des mots dans les corpus ;
- la méronymie (chapitre 8) : nous nous servons ici d'une base distributionnelle pour mettre à l'épreuve le caractère substituable des couples de méronymes et montrer quelles sont les raisons qui font que certains types de couples méronymiques ont tendance à être extraits par l'ADA alors que ce n'est pas le cas pour d'autres.

Le choix de nous focaliser sur ces relations a été motivé par le fait que ce sont des relations bien connues, qui ont été largement étudiées. De ce fait, nous nous appuyons par la suite sur les descriptions *traditionnelles* pour mettre en valeur les apports de l'ADA comme outil de description des relations lexicales.

Première partie

L'analyse distributionnelle automatique

Chapitre 1

Origine et principes théoriques

Sommaire

La plupart des études mettant en œuvre des méthodes dérivées de l'ADA font référence aux travaux du linguiste américain Zellig Harris (Harris, 1954, 1968; Harris *et al.*, 1989; Harris, 1991, ...). Ce dernier a en effet grandement contribué au développement de la méthode distributionnelle en la systématisant et en initiant son application à des textes de spécialité, ouvrant ainsi la voie à tout un champ de recherches sur la génération (semi-)automatique de thesaurus. Dans ce chapitre, nous donnons un aperçu du contexte intellectuel dans lequel l'AD est née ainsi que de la façon dont elle a évolué. Ce retour en arrière nous permet d'évoquer les fondements linguistiques qui sont à l'origine de l'AD et, ainsi, de mieux appréhender les problématiques rencontrées lors de la génération automatique de bases distributionnelles à partir de corpus.

Cette évolution est décrite en quatre points. À la section 1.1, nous évoquons l'*analyse en constituants immédiats*, qui est le principe phare de la

méthode d'AD proposée par Bloomfield. Nous montrons ensuite, à la section 1.2, comment cette méthodologie a été utilisée pour mettre au jour des classes sémantiques. À la section 1.3, nous décrivons l'approche développée par Harris *et al.* pour extraire des *schémas informationnels* à partir de textes de spécialité. Nous concluons, à la section 1.4, en évoquant une série de travaux menés entre 1961 et 1990 ayant en commun le fait d'exploiter – à différents niveaux – des méthodes d'analyse automatique de la langue dans le but de faire émerger automatiquement des classes de mots en s'appuyant sur le principe d'AD de Harris.

1.1 L'analyse en constituants immédiats

Le courant distributionnaliste a été initié aux États-Unis au début des années 30 par le linguiste américain Leonard Bloomfield. Il s'est répandu par la suite grâce à la diffusion de l'ouvrage fondateur qu'est *Language* (Bloomfield, 1935). Le distributionnalisme propose une façon d'aborder la langue inspirée de la pensée behavioriste, qui a pris naissance dans les années 20. Ainsi, de la même façon que la psychologie behavioriste propose d'étudier l'homme à travers l'observation de son comportement dans son milieu[1], le distributionnalisme s'appuie sur l'idée selon laquelle une unité linguistique se définit en fonction de ses contextes d'apparition dans les textes. De ce principe a découlé une méthode d'analyse se focalisant sur la description de la façon dont se combinent les unités que sont les phonèmes et les morphèmes.

La théorie distributionnelle s'inscrit dans un courant qui sera qualifié de *structuraliste* par écho au structuralisme saussurien qui se développe au même moment en Europe, suite à la parution du *Cours de linguistique générale* (de Saussure, 1916). De la même façon que les structuralistes européens, les distributionnalistes considèrent comme non arbitraire l'ordre des unités de la langue, qui est alors envisagée comme une structure composée d'entités entretenant des relations entre elles.

L'étude de la façon dont s'organisent ces unités s'appelle l'*analyse en constituants immédiats*. Bloomfield (1935, p. 161) évoque cette méthode d'analyse comme une décomposition successive des phrases en unités de différents niveaux, en partant de l'unité de plus haut niveau – la phrase – jusqu'aux morphèmes :

> Any English-speaking person who concerns himself with this mat-

[1] "Le comportement, c'est-à-dire la réaction de systèmes vivants aux facteurs du milieu, est le seul domaine qui puisse être étudié par la psychologie scientifique" (Köhler, 1929)

> ter, is sure to tell us that the *immediate constituents* of *Poor John ran away* are the two forms *poor John* and *ran away* ; that each of this is, in turn, a complex form ; that the immediate constituents of *ran away* are *ran* a morpheme, and *away*, a complex form, whose constituents are the morphemes *a-* and *way* ; and that the constituents of *poor John* are the morphemes *poor* and *John*.

Cette démarche s'appuie sur le principe de substitution, qui consiste à mettre sur un même niveau d'équivalence les unités qui peuvent apparaître dans un même contexte. Ainsi, si le corpus étudié contient des occurrences de *John ran away* ou de *Peter ran away*, alors on pourra en déduire que :

1. *poor John* forme une unité complexe (un *syntagme*),
2. cette unité est de la même nature que *John* et *Peter*.

Avec l'avènement de la grammaire générative, ces équivalences seront formulées à l'aide de règles de réécriture, qui explicitent les équivalences entre les différents constituants de la phrase.

La description de la façon dont se combinent ces unités revient en effet à construire la grammaire d'une langue, c'est-à-dire à faire l'inventaire 1) de ses classes d'unités élémentaires et 2) des différentes combinaisons dans lesquelles peuvent entrer ces classes. Toute combinaison qui n'apparaît pas dans le corpus étudié sera considérée comme *agrammaticale*. Cette démarche est dite *inductive* étant donné qu'elle permet d'inférer le système grammatical d'une langue sans faire appel à des connaissances extérieures (Bloomfield, 1970). C'est la raison pour laquelle l'AD – *via* l'analyse en constituants immédiats – a été en premier lieu mise en pratique dans le cadre de la description des nombreuses langues amérindiennes qui étaient encore indéchiffrées.

1.2 Faire émerger le sens

Dans la théorie distributionnelle, c'est donc la place que peut occuper une unité dans la phrase qui va déterminer sa catégorie d'appartenance. Une unité se retrouve alors définie par la somme des contextes dans lesquels elle apparait et les rapports qu'elle entretient avec les autres unités d'un mot ou d'une phrase. Autrement dit, une unité se définit par sa *distribution* :

> The distribution of an element will be understood as the sum of all its environments. An environment of an element A is an existing array of its co-occurrents, i.e. the other elements, each in a particular position, with which A occurs to yield an utterance. A's co-occurrents in a particular position are called its selection for that particular position. (Harris, 1954, p. 146)

Ce principe est central dans la théorie distributionnelle, dans le sens où, comme nous l'avons vu, il permet de rapprocher les unités qui apparaissent dans des contextes semblables afin de former des *classes distributionnelles*. Ces classes peuvent être de différentes natures. Nous avons vu à la section précédente qu'il pouvait s'agir de classes grammaticales comme celle des noms, des verbes, des adjectifs, etc. Mais il est également possible de mettre au jour des classes sémantiques.

La question du sens dans la théorie distributionnelle – et en particulier chez Harris – a fait l'objet de nombreuses discussions (Sahlgren, 2008). Dans le modèle distributionnel, le sens n'apparaît pas comme un critère pertinent pour la caractérisation des unités d'une langue. La raison en est que l'AD se veut une approche naïve, dans le sens où elle ne nécessite aucune connaissance *a priori* de la langue étudiée (cf. section 1.1). Par conséquent, le sens, qui "[recouvre] la somme des expériences communes d'un ensemble de locuteurs devant le [...] signe verbal" (Dubois, 1969) n'entre pas en ligne de compte dans la définition des classes distributionnelles d'une langue, ni dans la description de leur combinatoire.

Dans la méthode décrite par Harris, le sens apparaît comme un *construit* qui se dégage de la comparaison des contextes syntaxiques dans lesquels apparaissent les mots :

> The fact that, for example, not every adjective occurs with every noun can be used as a measure of meaning difference. (Harris, 1954, p. 156)

Dans cette conception, le sens d'un mot est envisagé en termes de sélections : un mot se définit par l'ensemble des contextes dans lesquels il apparaît. Cela revient à accepter la distribution d'un mot comme une représentation de son sens. De ce fait, le sens d'un mot est entièrement dépendant de l'usage qui en est fait dans le corpus étudié. Cette conception contextualiste a notamment été défendue par Martin Joos, Ludwig Wittgenstein et John Rupert Firth :

> Now the linguist's "meaning" of a morpheme [...] is by definition the set of conditional probabilities of its occurrence in context with all other morphemes. (Joos, 1950)[2]

> For a *large* class of cases – though not for all – in which we employ the word "meaning" it can be defined thus : the meaning of a word is its use in the language. (Wittgenstein, 1953)

> You shall know a word by the company it keeps. (Firth, 1957)

[2]Cité dans Osgood C., Suci G. et Tannenbaum P., *The measurement of meaning*, University of Illinois Press, 1967

Ainsi, de l'analyse des contextes syntaxiques des mots – comme les noms que peut modifier un adjectif, pour reprendre l'exemple de Harris – vont émerger des rapprochements qui traduisent les fonctionnements sémantiques en œuvre dans le corpus.

En effet, si le sens d'un mot se définit par ses contextes d'apparition, alors deux mots qui apparaissent dans des contextes similaires vont partager des éléments de sens. Et inversement, plus deux mots apparaissent dans des contextes différents, plus leurs sens sont éloignés :

> [...] if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Harris, 1954, p. 156)

> If A and B have some environments in common and some not (e.g. *oculist* and *lawyer*) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments. *id.*

C'est sur cette corrélation entre proximité distributionnelle et proximité de sens que s'appuient une partie des travaux en acquisition automatique de relations sémantiques.

Nous verrons par la suite, à la section 2.3, que les rapprochements générés par l'analyse peuvent être exploités pour faire apparaître des ensembles de mots présentant des similarités sémantiques. Ainsi, de la même façon qu'il est possible de dégager les classes grammaticales d'une langue et d'étudier la façon dont elles se combinent pour construire la grammaire de cette langue, l'AD permet d'extraire des classes sémantiques et de rendre compte de la façon dont elles sont structurées. Autrement dit, cette méthode donne la possibilité de construire ce que Habert et Zweigenbaum (2003) appellent des "grammaires sémantiques" qui rendent compte de la façon dont s'organise l'information dans les textes. Nous verrons que ce principe est largement exploité en traitement automatique des langues pour l'assistance à la construction d'ontologies.

1.3 La méthodologie harrissienne

Dans cette section, nous évoquons les raisons qui ont amené Zellig Harris et ses collègues travaillant sur l'AD à orienter leurs travaux en direction des textes de spécialité. En effet, le développement du modèle distributionnel est intimement lié à la théorie harrissienne des sous-langages, qui distingue les

textes qui relèvent de domaines spécialisés de ceux qui relèvent de la *langue générale*.

La nécessité d'opérer une telle dichotomie trouve sa légitimité dans le fait que les textes produits dans le contexte de pratiques professionnelles possèdent des propriétés différentes des textes qui appartiennent à la langue non spécialisée. Nous nous focalisons par la suite sur les propriétés linguistiques de ces textes, et plus particulièrement leurs propriétés lexico-syntaxiques (section 1.3.1). Nous décrivons ensuite la méthode développée par Harris *et al.* (1989) pour extraire des schémas informationnels à partir de ces textes (section 1.3.2).

1.3.1 La théorie des sous-langages

Les textes de spécialité se distinguent des textes qui relèvent de la langue générale du fait que les contraintes sélectionnelles qui pèsent sur les mots qui les composent (les *termes*) sont de type binaire alors que, dans la langue *générale*, on observe un mode de fonctionnement de type probabiliste (que Dubois (1969) met en parallèle avec le modèle markovien). Ainsi, dans la langue générale, un verbe comme *dormir* prend typiquement en position objet des agents animés comme *bébé* ou *chat*. On dit que les mots de ce type ont une *vraisemblance d'occurrence* (Buvet et Grezka, 2009) élevée en position sujet de *dormir*. Cependant, l'ensemble des noms qui peuvent occuper cette position est virtuellement ouvert à l'ensemble des noms du lexique. Cela est dû au fait que la langue générale permet l'emploi de figures de style comme la métonymie (*la ville dort*) ou de métaphores (*l'eau dort*). De fait, même si certains mots ont tendance à apparaître plus souvent que d'autres dans certaines positions, la probabilité qu'un mot apparaisse dans une position donnée n'est jamais nulle – si tant est que sa catégorie grammaticale le lui permet :

> The flexibility of language that enables it to accommodate nonsense, fairy tales, untruths, and so on, leaves it to the speaker to choose whichever word seems suitable as long as they are assembled into an understandably grammatical sentence. (Sager et Ngô Thanh, 2002, p. 91)

Ce type de fonctionnement ne s'observe pas dans des textes relevant d'un domaine de spécialité. Les langues de spécialité se distinguent en effet de la langue générale sur deux points. D'une part, les restrictions qui pèsent sur les contraintes sélectionnelles sont beaucoup plus fortes. Ainsi, Habert (1998) fait remarquer que dans son corpus d'étude (le corpus Menelas, constitué de manuels, de comptes-rendus d'hospitalisation et de lettres adressées à des

médecins portant sur le thème des maladies coronariennes (Zweigenbaum et Consortium MENELAS, 1994)), le verbe *dilater* se construit de façon systématique avec un terme appartenant à la classe des médecins en position sujet et un terme appartenant à la classe des artères (*artère coronaire*, *artère circonflexe*, etc.) en position objet.

D'autre part, l'influence du domaine de spécialité sur la langue peut entraîner la modification de la structure argumentale des mots. Habert (1998) cite l'exemple du verbe *sucrer*, qui, lorsqu'il est employé dans des textes relevant du domaine de l'œnologie, s'emploie de façon intransitive. La raison en est que, dans le processus de vinification, le sucre ne peut s'ajouter qu'au moût[3]. Il devient donc superflu d'exprimer l'argument : **sucrer le moût* constitue un énoncé agrammatical.

La façon dont se combinent les mots dans un texte de spécialité reflète ainsi les pratiques du domaine. Les textes de spécialité peuvent donc se caractériser par les contraintes sélectionnelles qu'ils imposent à leurs opérateurs. Kittredge (2003) montre que la nature des arguments du verbe *to sweep* varie selon qu'il est employé dans des compte-rendus de matchs de base-ball (1) ou dans des articles d'entomologie (2) :

(1) <team_1> sweep <team_2> <string_of_scores>
The Redbirds swept Laval 10–4 and 12–5.

(2) <mecopteran_insect> sweep <body_part> over <vegetation>
Males swept hind legs over vegetation.

De par la "coïncidence entre la structure de l'information et la structure des sous-langues de la science" (Martinot, 2007, p. 3), l'étude des contraintes qui pèsent sur les sélections opérateur–argument peut donc constituer un moyen d'accéder à la "grammaire sémantique" des textes de spécialité et, par conséquent, permettre d'expliciter la façon dont s'organise l'information dans ces domaines :

> The lexical classes and the hierarchical relations between the classes usually reflect the accepted taxonomy which the specialized field of knowledge imposes on the objects of its limited domain of discourse. And the combinations of lexical classes which are permissible in the sentences of the specialized texts reflect the conceivable relations between these objects (regardless of the truth or falsity). (Kittredge, 1982, p. 112)

[3]La définition que donne le TLFi du verbe *sucrer* dans cette acception est la suivante : "Addition de sucre au moût de raisin, avant la fermentation, pour élever sa teneur en alcool ou pour élaborer certains vins effervescents".

L'analyse des structures argumentales des mots dans des corpus spécialisés permet, de ce fait, de faire émerger des classes d'entités dont l'interprétation est ensuite confiée à un expert du domaine. C'est sur ce principe que s'appuient les travaux qui exploitent l'AD pour assister la construction d'ontologies.

Dans *The form of information in science*, Harris et son équipe mettent au point une méthode qui s'appuie sur l'AD pour rendre compte de la façon dont s'organisent les connaissances dans les textes de spécialité. Nous la décrivons ici dans ses grandes lignes.

1.3.2 *The Form of Information in Science*

Nous rapportons ici la méthodologie développée par Harris et collègues dans *The Form of Information in Science : Analysis of an immunology sublanguage*, paru en 1989. Cet ouvrage présente en détails une méthode distributionnelle visant à analyser les fonctionnements lexico-syntaxiques observables dans des textes de spécialité pour en extraire des grammaires. Ces grammaires prennent la forme de schémas informationnels indiquant les classes de mots qu'un opérateur peut prendre comme arguments. Cette démarche repose sur l'hypothèse selon laquelle les contraintes lexico-syntaxiques qui régissent les textes relevant d'un domaine donné sont révélatrices des pratiques du domaine en question. Nous présentons ici cette méthodologie dans ses grandes lignes.

Le corpus sur lequel travaillent Harris et son équipe est constitué d'articles relevant du domaine de l'immunologie. Le principe de base qu'ils mettent en œuvre reste celui de l'AD telle que décrit à la section 1.2 :

> The words are identified not by their meanings but by the combinations into which they enter in respect to other words, within each sentence of the science material. (Harris *et al.*, 1989, p. 1)

> In principle, word classes in a closed corpus of texts are established by characterizing each word-occurrence by its "co-occurrents", i.e. the words to which it has a grammatical relation in a sentence, and then putting into one class those word-occurrences which have the same co-occurrents, or nearly the same. (*id.*, p. 29)

L'analyse des relations récurrentes verbe/sujet et verbe/objet fait apparaître des classes d'entités, comme la classe des antigènes (*antigen*, *pneummococcus*, *influenza virus*, etc.), notée G, dont les membres cooccurrent fréquemment avec la classe J, qui regroupe toutes les entités qui renvoient – en contexte – à l'action d'injecter (*inject*, *injection*, *administered*, etc.). Cette formalisation

permet de mettre au jour la “formule” GJ dans les phrases (3) et (4) :

(3) Parathyphoid bacteria was injected on one side.

(4) These animals were challenged with tetanus toxoid.

Il est à noter que ces deux phrases contiennent la même formule alors que G apparaît en position sujet de J dans (1) et en position objet dans (2). Les auteurs ont en effet appliqué aux phrases du corpus un ensemble de modifications. L'une d'elles consiste à ramener sous une même forme canonique les structures passives et actives, ou encore les nominalisations comme *antigen injection* (cette procédure renvoie à la notion de transformation, centrale dans la théorie harrissienne).

Les formules peuvent couvrir des phrases entières, comme c'est le cas pour la phrase (5), qui équivaut à la formule AVT, où A renvoie à la classe des anticorps, V à la classe des opérateurs qui expriment une relation d'inclusion (typiquement *is present in*) et T à celle des tissus biologiques.

(5) Antibody is contained in suspensions.

Cette façon de ramener les phrases à des formules permet, à terme, de dresser l'inventaire des combinaisons de classes d'entités qui sont *possibles* pour un corpus et un domaine donnés, de la même façon que la grammaire d'une langue distingue les combinaisons de classes syntaxiques qui sont *grammaticales* de celles qui ne le sont pas. Cette méthode possède en outre l'avantage de pouvoir s'appliquer sur tout type de texte de spécialité et de ne pas être dépendante d'une seule langue (Harris et ses collègues l'ont également testée sur le français).

Les auteurs de l'étude évoquent comme perspective l'automatisation de cette procédure :

> At least in part, the methods could be carried out in computer programs applied to the articles as published, without pre- or post-editing. (Harris *et al.*, 1989, p. 1)

L'automatisation des méthodes distributionnelles représentait en effet un enjeu de taille étant donné que l'extraction d'informations lexico-syntaxiques facilite considérablement le travail de l'expert lors de la délimitation des classes conceptuelles d'un domaine et des relations qu'elles entretiennent. Une des limites à l'application à grande échelle de la méthodologie harrissienne telle qu'elle est décrite dans *The Form of Information in Science* est l'importance de l'intervention humaine dans le processus d'annotation syntaxique et de transformation des phrases.

1.4 Automatiser l'AD

Nous avons montré jusque-là comment Harris et ses collègues ont utilisé les principes de l'analyse distributionelle pour mettre en place une méthodologie permettant d'extraire des classes sémantiques à partir de textes de spécialité. Ces travaux ont joué un rôle essentiel dans la diffusion de la méthode d'AD. En effet, ils proposent un cadre d'analyse défini (les textes de spécialité) ainsi qu'une méthode dont l'automatisation permettrait de fournir une réponse aux besoins naissants en extraction d'informations sémantiques.

Une des raisons pour lesquelles le domaine du TAL s'est emparé de ce principe est qu'il repose sur une analyse systématique des données textuelles, le travail interprétatif étant extérieur à l'étape de définition des classes. Il est donc particulièrement commode à automatiser. Sa mise en œuvre ne requiert qu'un corpus analysé syntaxiquement[4] ainsi qu'un programme qui procède à l'analyse distributionnelle elle-même. Toutefois, la difficulté de disposer d'un analyseur syntaxique opérationnel a constitué un frein à la popularisation du modèle distributionnel : les premiers travaux qui ont porté sur l'automatisation de l'AD montrent qu'une part importante des traitements apportés au corpus était encore effectués manuellement. L'arrivée d'analyseurs syntaxiques dits "robustes" dans les années 90 a marqué un tournant dans le développement de l'ADA. Une analyse syntaxique automatique permet en effet d'échapper à la fastidieuse tâche d'annotation manuelle et, par conséquent, de pouvoir appliquer la méthode distributionnelle sur les corpus de plus en plus volumineux qui ont été disponibles par la suite. En permettant les analyses de grands volumes de textes, les analyseurs ont donc entrainé un regain d'intérêt pour la méthode distributionnelle, notamment dans le domaine de l'ingénierie des connaissances.

Nous décrivons ici quelques-uns des travaux qui ont été menés durant la période qui a précédé l'apparition des analyseurs, afin d'illustrer les différentes étapes qui ont mené au développement du modèle d'ADA tel qu'il s'est répandu par la suite.

Harper (1961, 1965) À notre connaissance, les premiers travaux portant sur l'implémentation d'une analyse inspirée de l'approche distributionnelle sont ceux de Kenneth Eugene Harper, chercheur à la RAND Corporation. Contexte historique oblige, ses travaux portent sur des corpus de textes rédigés en russe issus du domaine de la physique. Dans Harper (1961), l'annotation syntaxique du corpus a été menée de façon semi-automatique, à l'aide

[4]Nous verrons que la méthode distributionnelle peut aussi bien opérer sur des corpus qui ne sont pas analysés syntaxiquement.

d'un lexique-grammaire. L'auteur utilise ensuite un ensemble de mots annotés manuellement – les noms d'animés, les verbes de mouvement, les adjectifs de couleur, etc. – pour former des classes distributionnelles. Sa méthode lui permet d'extraire des classes comme celle des verbes qui ne prennent que des noms désignant des animés en position sujet, ou des particules en position objet. Ainsi, bien que l'auteur ne fasse aucune référence aux travaux des distributionnalistes, on reconnaît là le type de démarche qui sera employé plus tard par Harris *et al.* (1989).

Dans Harper (1965), l'idée d'utiliser des mots annotés sémantiquement est abandonnée au profit d'une approche non supervisée :

> Rather than assign syntactic or "semantic" codes on this basis, however, we might proceed statistically : a computer program will be devised for counting the number (and considering the frequency) of certain syntactically equivalent dependents that each pair of verbs has in common. The degree of alikeness is represented by a number (the association coefficient), and through these numbers classes of words are distinguished. (p. 9)

On retrouve l'idée selon laquelle l'analyse de la distribution des mots du corpus peut permettre à elle seule de faire émerger des classes. Dans cette étude, l'auteur cherche à vérifier l'hypothèse d'une corrélation entre proximité distributionnelle et proximité sémantique à l'aide d'une tâche de classification : quarante noms russes sont choisis dans le corpus utilisé dans Harper (1961) et sont manuellement répartis dans 11 groupes en fonction de leur sens. La démarche consiste à comparer entre elles les distributions de ces quarante mots[5] et de voir s'il y a un recouvrement entre les groupes formés *a priori* et ceux qui émergent de l'analyse automatique. Les résultats montrent que les regroupements extraits vont au-delà des ensembles produits manuellement :

> Groups [...] which include the names of chemical mixtures, classes of elements, individual elements, and components of elements, may be taken together semantically as a single sub-class of "object nouns". The physicist tends to say the same things about all the nouns in this group. (Harper, 1965, p. 17)

Ainsi, malgré les limites du protocole mis en place[6], l'étude valide l'hypothèse distributionnelle en mettant toutefois en évidence un décalage entre la

[5]Pour un mot donné, sont pris en compte les adjectifs et les noms avec lesquels il entretient une relation de dépendance.

[6]Que l'auteur reconnaît bien volontiers : "A few aditionnal remarks may be made about the procedure itself, which may be likened to deep-sea fishing with a tea strainer full of holes."

granularité des classes de mots qu'il est possible de définir *a priori* et celle qui émerge d'une analyse de corpus.

En conclusion, malgré le peu d'écho qu'ils ont reçu dans la communauté scientifique, les travaux de Harper apparaissent comme pionniers dans le domaine de l'étude des méthodes distributionnelles. Ils évoquent en effet, dès le milieu des années 60, des problématiques comme l'interprétation des résultats fournis par l'AD, l'influence du type de corpus ou encore le problème de la polysémie pour la classification, soit autant de questions qui constituent encore aujourd'hui des sujets de recherche.

Rubenstein et Goodenough (1965) L'étude de Rubenstein & Goodenough consiste à appliquer automatiquement la théorie distributionnelle afin d'éprouver son potentiel à repérer des mots qui entretiennent une relation de synonymie. Le protocole mis en place implique la constitution d'un jeu de 65 paires de noms auxquelles des locuteurs ont associé un score de similarité entre 0 – si les mots sont éloignés – et 4 – s'ils sont proches : *gem/jewel* (3,94), *magician/oracle* (1,82), *rooster/voyage* (0,04), etc. Ce jeu de couples est comparé aux rapprochements générés par une approche automatique. Cette dernière est appliquée sur un corpus de 4800 phrases entièrement rédigé pour l'occasion. Le principe est de calculer un score de similarité entre les mots du jeu de couples basé sur la comparaison de leurs contextes d'apparition dans le corpus. Le contexte est ici défini comme l'ensemble des mots qui apparaissent dans la même phrase que le mot cible. Les résultats montrent une corrélation entre les scores de similarité définis par les locuteurs et celui qui a été calculé à partir du corpus. Cette corrélation est d'autant plus forte que les deux mots partagent un nombre important de contextes.

Cette étude est intéressante à plusieurs niveaux. Elle se distingue dans un premier temps par le fait qu'elle se situe en dehors du cadre des textes de spécialité, qui constituent le cadre dans lequel a été développée l'AD harrissienne. De la même façon, en posant une définition du contexte basée sur la simple cooccurrence, elle réinterprète la théorie distributionnelle, qui, originellement, pose une équivalence entre les mots qui partagent les mêmes opérateurs ou opérandes. Nous verrons que cette alternative a connu par la suite un succès comparable aux approches qui s'appuient sur une conception syntaxique du contexte (section 2.1.2). Enfin, cette étude constitue la première tentative d'utiliser des méthodes issues de la psycholinguistique pour évaluer des données acquises à partir de corpus. Il est à noter que la ressource qui a été générée dans le cadre de ces travaux sera par ailleurs réutilisée dans de nombreuses études portant sur les mesures de similarité sémantique (Resnik, 1995; Budanitsky et Hirst, 2001; Banerjee et Pedersen, 2003; Baroni et

Lenci, 2010).

Hirschman *et al.* (1975) Cette étude s'appuie sur les travaux de Naomi Sager, qui a montré, dans la lignée de Harris (1968), que le discours scientifique possédait des propriétés qui le distinguaient de la langue générale et que ces propriétés pouvaient être exploitées pour mettre au jour des classes de verbes et de noms :

> An earlier study indicated that the parts of the sentence carrying the scientific information fell into a small number of patterns, called information formats : certain groups of verbs occurred only with certain other groups of nouns as subjects and objects. Furthermore, these groups correlated closely with the intuitive semantic classes in the field. This suggested that word classes pertinent to the informational structure of the sentences could be obtained from an analysis of the subject-verb-object co-occurrence statistics. (Hirschman *et al.*, 1975, p. 39)

Le but est donc ici de mettre au jour des classes de mots afin de reconstituer les structures informationnelles exprimées dans les textes de spécialité. Les auteurs s'inscrivent donc dans le domaine naissant de la génération automatique de thésaurus.

La démarche s'appuie sur une analyse syntaxique manuelle de 400 phrases issues d'un corpus de 6 articles de pharmacologie. Une fois l'analyse effectuée, un programme génère des couples entre un opérateur et ses arguments[7] et rapproche les couples d'opérateurs qui partagent les mêmes arguments – dans les mêmes positions – et les couples d'arguments qui s'emploient avec les mêmes opérateurs. Les opérateurs et arguments sont agrégés pour former des classes (classes des muscles ou des enzymes pour les arguments, classes des mouvements ou des changements d'état pour les opérateurs) qui sont ensuite automatiquement comparées avec des classes construites manuellement par des experts. L'évaluation des classes générées montre que :

- 10 des 11 classes mises au jour par l'expert ont été extraites automatiquement ;
- moins de 9 % des 43 opérateurs et arguments pris en compte apparaissent dans la mauvaise classe.

Les résultats de cette étude s'avèrent donc particulièrement concluants. Ils marquent le point de départ de tout une série de travaux portant sur l'extraction de classes à partir de corpus médicaux.

[7]Par exemple, l'analyse du SN *potassium loss from heart caused by CG* permettra de générer les couples *opérateur–argument 1* <cause,CG> et <lose,heart> et les couples *opérateur–argument 2* <cause,lose> et <lose,potassium>.

Hindle (1990) Cette étude s'inscrit explicitement dans la lignée des travaux de Hirschman *et al.* Elle s'en distingue toutefois par le fait que le corpus utilisé a été annoté automatiquement, ce qui permet d'éprouver la théorie distributionnelle sur un corpus de grande taille – relativement à ceux qui ont été utilisés jusque-là. Ce corpus, constitué d'articles de presse, compte en effet 6 millions de mots. L'étude se focalise sur le rapprochement des noms en fonction des verbes dont ils sont le sujet ou l'objet.

L'application de l'AD à grande échelle permet de mettre au jour des problématiques qui n'apparaissaient pas jusque-là. Par exemple, alors que 9 des 10 mots qui ont la distribution la plus similaire à celle de *boat* présentent une certaine homogénéité thématique – ils renvoient tous à des moyens de transport (*ship*, *plane*, *bus*, etc.) –, les mots qui partagent le plus de contextes avec *table* sont de nature beaucoup plus hétérogène (*farm*, *river*, *town*, etc.). Hindle semble ici payer les conséquences de son choix de corpus : on peut remarquer qu'au moins deux acceptions de *table* se dégagent à travers les contextes (le meuble et la structure de données). De plus, il apparaît que les mots qui ont été rapprochés de *table*, sous leur hétérogénéité apparente, partagent – pour la plupart – la propriété de renvoyer à des noms pouvant être interprétés comme des lieux : *farm*, *scene*, *America*, *town*, *hospital*, etc. Les contextes qui ont permis de rapprocher ces mots de *table* sont des verbes exprimant le déplacement : *come to* _, *go to* _, *return to* _, *leave*, etc. Cela signifie que le mot *table*, dans le corpus utilisé par Hindle, renvoie principalement à une entité localisée dans l'espace qui peut être soit la cible soit la source d'un déplacement.

Ces travaux constituent donc une étude pionnière sur le sujet, puisque grâce à la mise en place d'un protocole entièrement non supervisé, ils touchent du doigt les problématiques liées à l'utilisation de l'ADA sur des corpus non spécialisés. De plus, ils marquent un tournant dans l'histoire du développement de cette méthode, dans le sens où l'utilisation d'un corpus de plusieurs millions de mots constitue un changement d'échelle radical par rapport aux études que nous avons décrites jusque-là.

Avec le développement d'analyseurs syntaxiques robustes et la disponibilité de quantités de textes au format numérique toujours plus importantes, l'ADA a par la suite fait l'objet d'un nombre grandissant d'études qui seront présentées dans la section suivante. Dans le même temps, elle a été déclinée en une variété de modèles qui se distinguent en fonction de la configuration de paramètres adoptée[8] (utilisation de données analysées syntaxiquement, type de mesure de similarité, etc.). Dans le chapitre suivant, nous faisons

[8]Nous faisons un panorama des modèles actuels à la section 3.1.

le point sur les différents paramètres qui entrent en jeu lors de la mise en œuvre d'une analyse automatique de corpus ainsi que sur l'influence que ces paramètres peuvent avoir sur les résultats obtenus.

Chapitre 2

Mise en œuvre

Sommaire

Nous avons vu jusqu'ici que le principe sur lequel repose l'approche distributionnelle est que le sens émerge de l'analyse des rapports qu'entretiennent les mots entre eux. Pour Grefenstette (1994a), ces rapports – ou *affinités* – se distinguent en trois types :

- une affinité de premier ordre, qui opère entre un mot et les unités qui apparaissent dans son contexte immédiat (relation de type syntagmatique) ;
- une affinité de deuxième ordre entre les mots qui partagent les mêmes contextes d'apparition sans forcément co-occurrer (relation de type paradigmatique) ;

– une affinité de troisième ordre entre les sous-groupes de mots distributionnellement similaires (extraction des *classes*).

Ces trois affinités correspondent aux procédures nécessaires à la mise au jour de classes sémantiques. Ces étapes se déroulent de façon incrémentale : les affinités de deuxième ordre se calculent en fonction des affinités de premier ordre, de la même façon que les classes – affinités de troisième ordre – se construisent à partir de l'analyse des relations de deuxième ordre.

L'implémentation de chacune de ces étapes implique des choix. Elle peut en effet varier en fonction de plusieurs facteurs. Faute de méthode d'évaluation stable des résultats fournis par les méthodes distributionnelles, l'influence de ces choix sur la nature des rapprochements générés est souvent difficile à apprécier. C'est la raison pour laquelle les réglages sont fréquemment adoptés de façon empirique, "par tâtonnement".

Ce chapitre se compose de trois sections calquées sur les ordres définis par Grefenstette (1994a). Nous y présentons les différents paramètres sur lesquels il est possible d'influer lorsque l'on met en œuvre une ADA de corpus.

2.1 Extraire les contextes

L'AD s'appuie sur le principe selon lequel deux mots qui apparaissent dans les mêmes contextes ont des chances d'entretenir une relation de sens. Ainsi, c'est la somme des contextes dans lesquels apparaît un mot qui va définir ce mot et servir de point de comparaison lors du rapprochement avec les autres mots du corpus. La définition de ce que l'on considère comme étant le contexte d'un mot revêt donc une importance capitale.

Dans cette section, nous passons en revue les différentes étapes qui composent la démarche d'extraction des contextes.

2.1.1 Prétraitement du corpus

Une première étape dans l'extraction des contextes consiste à appliquer au corpus une série de traitements comme la tokenization, la lemmatisation ou l'étiquetage morphosyntaxique (Habert, 2005). Ces manipulations visent à obtenir une représentation du texte favorable à la modélisation des contextes pour l'ADA.

L'étape préalable consiste à segmenter la chaîne de caractères que constitue le texte en une série de mots (*tokenization*). Le traitement qui est le plus fréquemment réalisé ensuite est l'annotation morphosyntaxique, qui consiste à associer à chaque mot sa catégorie grammaticale et son lemme. De nombreux programmes permettent de réaliser une telle opération. L'un des plus

répandus, TreeTagger[1] (Schmid, 1994), construit un modèle de langue par apprentissage sur un corpus annoté. Il s'appuie sur un lexique ainsi que sur des arbres de décision pour calculer la probabilité d'apparition d'une étiquette grammaticale donnée en fonction du type d'étiquettes qui figurent dans son contexte. Après lemmatisation, toutes les formes fléchies des verbes, noms, adjectifs et déterminants sont regroupées sous une même forme, le lemme. Cela a une influence sur les résultats de l'ADA étant donné que cela augmente considérablement le grain de l'analyse : tous les contextes des formes morphologiquement dérivées des lemmes sont alors confondus. De fait, cette procédure a l'avantage d'atténuer le phénomène de sparsité (*sparseness*), en plus de permettre de réduire la complexité calculatoire lors du rapprochement des couples de mots.

À l'heure actuelle, l'étiquetage morphosyntaxique est une tâche relativement bien maîtrisée : Giesbrecht et Evert (2009) montrent qu'en moyenne, sur des textes en anglais, les performances des cinq étiqueteurs les plus répandus[2] varient entre 98 % de mots correctement analysés – sur du texte journalistique – et 93 % – sur du texte issu du Web. L'impact de l'étiquetage morphosyntaxique sur la qualité des résultats finaux reste toutefois incertain. Ainsi, Pekar (2004) ne remarque aucune différence de performance significative dans une tâche de classification. Sahlgren (2006) évoque les travaux de Wiemer-Hastings et Zipitria (2001) et Widdows (2003), qui rapportent des résultats mitigés montrant que dans certains cas, le fait de préciser la partie du discours des mots du corpus pouvait entraîner une baisse de performance des ressources générées dans des tâches comme l'extraction de synonymes.

L'annotation morphosyntaxique peut servir de support à d'autres annotations de plus haut niveau, de type syntaxique ou sémantique. L'analyse syntaxique de surface (ou *chunking*) consiste en un découpage des phrases en syntagmes. Elle ne cherche donc pas à analyser la structure syntaxique de la phrase dans son intégralité, c'est une analyse partielle (elle peut toutefois constituer la première étape d'une analyse syntaxique plus poussée). Cette procédure est utilisée dans le cadre de l'extraction terminologique (Bourigault, 1993) ou pour l'identification d'unités lexicales complexes – *passer à tabac*, *pomme de terre* – qui seront alors considérées comme des unités à part entière lors du calcul distributionnel. L'étiquetage sémantique consiste à annoter les mots d'un texte afin de lever les ambiguïtés sémantiques (Ide et Véronis, 1998). Cette couche est rarement appliquée étant donné que la désambiguïsation automatique en est encore à l'état de sujet de recherche

[1] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger\index{TreeTagger}/`

[2] TreeTagger, TnT, SVMTagger, Stanford tagger et Apache UIMA Tagger.

et que la désambiguïsation manuelle de tout un corpus est inenvisageable. Elle constitue cependant un enjeu dans le sens où elle permettrait de capter la distribution de différentes acceptions d'une unité polysémique. Le fait de traiter les deux occurrences d'un mot polysémique comme une seule et même unité lexicale donne en effet naissance à des entités hétérogènes qui cumulent les distributions des différentes acceptions qui les composent, ce qui a pour conséquence de gêner l'interprétation des rapprochements générés.

2.1.2 Contextes syntaxiques vs fenêtres de mots

L'étape qui suit consiste à définir les conditions qui font qu'un mot peut être considéré comme appartenant au contexte d'un mot cible donné[3]. Nous avons vu à la section 1.4 que, dès les premières implémentations de l'ADA, deux façons de considérer le contexte d'un mot ont apparu. Aujourd'hui encore, on peut ainsi distinguer ces deux tendances :

- certains modèles adoptent une conception syntaxique du contexte ;
- d'autres définissent le contexte comme relevant d'une relation de cooccurrence simple.

Nous décrivons ici ces deux approches avant d'en faire un comparatif.

2.1.2.1 Contextes syntaxiques

Les approches syntaxiques envisagent le contexte d'un mot comme l'ensemble des mots qui entretiennent avec lui une relation de dépendance syntaxique (Hindle, 1990; Grefenstette, 1992b; Lin, 1998a; Bourigault, 2002; Curran, 2004; Padó et Lapata, 2007; Baroni et Lenci, 2010; Henestroza Anguiano et Denis, 2011, etc.) Cette conception du contexte se conforme à celle de l'AD harrissienne :

> [...] the structural property is not merely co-occurrence, or even frequent co-occurrence, but rather dependence of a word on a set [...] (Harris, 1991, p. 332), cité par Habert et Zweigenbaum (2002)

Elle implique une étape de prétraitement supplémentaire qui consiste à fournir au système une analyse préalable des relations syntaxiques qui opèrent entre les constituants des phrases du corpus. Étant donnée l'absence d'analyseur syntaxique opérationnel, les premières tentatives d'automatiser l'AD ont nécessité une annotation manuelle longue et fastidieuse. Cette contrainte a freiné l'application de l'AD sur des corpus de grande taille. Aujourd'hui,

[3]Nous reprenons par la suite le terme *mot cible* (*target word*) pour désigner le mot dont on cherche à extraire les contextes.

il est possible de disposer d'analyseurs robustes qui permettent de fournir des analyses complètes de phrases, et ce avec une fiabilité satisfaisante (la campagne d'évaluation EASY (Paroubek *et al.*, 2006) a permis de montrer que les meilleurs analyseurs obtenaient une mesure F de 0,92 pour la relation sujet sur certains types de corpus). L'analyseur génère ainsi des arbres syntaxiques qui permettent de relier chaque mot de la phrase à ses dépendants, lesquels serviront de contextes. Ces derniers sont traditionnellement extraits sous la forme de triplets <*mot1,* RELATION*, mot2*> (<*rouler,* SUJ*, voiture*>). Nous verrons à la section 3.1.2, que des travaux comme ceux de Baroni et Lenci (2010) proposent des alternatives à ce modèle.

Nous montrons au tableau 2.1 les contextes du nom *couvent* extraits par une analyse syntaxique de la phrase suivante :

(1) En 1525, les paysans détruisirent le couvent des carmélites et pillèrent le monastère des dominicains[4].

On peut voir que la fonction complément portée par *carmélite* apparaît sous la forme de la préposition *de*. Cette façon de modéliser les syntagmes prépositionnels qui sont compléments ou modifieurs est celle qui a été adoptée dans la chaîne Syntex-Upéry (section 3.2). Cette information est importante dans la mesure où les prépositions donnent un indice de la relation entretenue par un verbe ou un nom et leurs compléments. Ainsi, dans un corpus constitué d'articles de Wikipédia[5], les contextes de *bataille* peuvent renvoyer à des entités différentes selon la préposition avec laquelle le mot s'emploie :

- des lieux (*Constantinople*, *Valmy*, *Normandie*) avec les prépositions *de*, *en*, *à sur* et *au large de* ;
- des individus ou des groupes d'individus (*soldat*, *Russes*, *armée*) avec les prépositions *contre* et *entre* ;
- des événements (*guerre mondiale*, *guerre*) avec la préposition *durant* ;
- des concepts (contrôle, liberté, indépendance) ou des lieux (*Terre du Milieu*, *Dune*, *Terre*) avec la préposition *pour*.

Dans cette optique, une conception encore plus lexicalisée du contexte peut être adoptée. Par exemple, le modèle *LexDM* de Baroni et Lenci (2010) lexicalise toutes les relations syntaxiques. Ainsi, dans le triplet <*rejoindre,* OBJ*, couvent*>, la relation objet prend place dans un triplet qui fait apparaître le verbe, son objet, mais aussi son sujet (cf. tableau 2.1).

[4] `http://fr.wikipedia.org/wiki/Iena`

[5] Ce corpus – désormais *corpus Wikipédia* – a permis de générer une des ressources distributionnelles – les *voisins de Wikipédia* – que nous utilisons dans la suite de ce travail. C'est de cette ressource que sont extraits la plupart des exemples qui figurent dans ce chapitre.

Modèle syntaxique	
Semi-lexicalisé	<*détruire,* OBJ*, couvent*>, <*couvent,* DE*, carmélite*>
Lexicalisé	<*paysan,* DÉTRUIRE*, couvent*>, <*couvent,* DE*, carmélite*>

TAB. 2.1 – Contextes syntaxiques du nom *couvent* extraits de la phrase (1).

L'avantage d'avoir des contextes de statuts différents est que cela permet d'opérer une sélection parmi ceux qui serviront pour le calcul de proximité distributionnelle et les autres. Ainsi, nous avons vu que Hindle (1990) définissait la distribution des noms comme l'ensemble des verbes dont ils sont le sujet ou l'objet. Dagan *et al.* (1997) et Weeds (2003) font, eux, le choix d'exclure la relation sujet. La question de savoir quel type de relation conserver dans un modèle distributionnel syntaxique et quelles en sont les conséquences sur les rapprochements générés n'a pas encore de réponse claire. Van der Plas (2008) a montré que le fait de rapprocher les noms sur la base d'une seule et même relation syntaxique favorisait – légèrement – la mise au jour de couples comme les co-hyponymes ou les hypo/hyperonymes. Par exemple, les couples d'hypo/hyperonymes ont tendance à être mieux captés lorsque la relation qui a servi de contexte est la relation objet. Toutefois, ses résultats montrent clairement que c'est la combinaison de plusieurs relations qui permet globalement d'extraire les couples les plus pertinents (les rapprochements sont évalués par une comparaison avec une ressource de référence). Cela confirme les résultats de Lin (1998a); Padó et Lapata (2007).

2.1.2.2 Approches à fenêtres de mots

Nous appelons "approches à fenêtres de mots" les implémentations de l'ADA qui envisagent le contexte d'un mot comme étant l'ensemble des mots qui cooccurrent avec lui dans une fenêtre textuelle définie *a priori* (Rubenstein et Goodenough, 1965; Deerwester *et al.*, 1990; Lund et Burgess, 1996; Schütze, 1998; Turney, 2001; Sahlgren, 2006; Ferret, 2010, etc.). Contrairement à l'approche syntaxique, ce rapport de cooccurrence est purement positionnel, il n'est pas contraint par l'impératif pour les cooccurrents d'entretenir une relation fonctionnelle.

Dans les approches syntaxiques, un mot ne peut avoir de contexte au delà du cadre de la phrase (on n'envisage pas de relation syntaxique interphrastique). Dans les approches à fenêtres de mots, l'absence de ce critère syntaxique offre la possibilité de paramétrer la taille ainsi que la nature de la fenêtre. Ses limites peuvent être soit :

– celles d'unités textuelles (phrase, paragraphe, document) ;

– définies comme l'intervalle de $\pm$ n mots à droite et/ou à gauche du mot dont on cherche à extraire le contexte.

La taille de la fenêtre utilisée peut varier de quelques mots ($3+3$ chez Dagan *et al.* (1993)[6]) à plusieurs dizaines ($50+50$ chez Gale *et al.* (1994)). La plupart des approches s'appuient sur un étiquetage morphosyntaxique préalable pour filtrer certaines catégories grammaticales (classiquement, les mots *pleins* que sont les noms, verbes et adjectifs). Plus la fenêtre est réduite, plus les contextes extraits ont des chances de ressembler aux types de contextes récupérés par l'approche syntaxique. Une fenêtre plus étendue a tendance à fournir des contextes plus lâches qui sont liés au mot cible par une association de type thématique (Patel *et al.*, 1997; Habert et Zweigenbaum, 2002; Rapp, 2002)[7].

Un autre critère qu'il est possible de paramétrer est la direction de la fenêtre. Selon les informations que l'on veut récupérer, la fenêtre peut se placer d'un seul côté du mot cible – à gauche pour récupérer les sujets d'un verbe, à droite pour ses objets – ou couvrir simultanément les contextes gauche et droit.

Par défaut, l'approche à fenêtre de mots est *non structurée* (Evert *et al.*, 2010) : la seule information qui lie le mot cible à un mot appartenant à son contexte est qu'ils cooccurrent. Le lien de cooccurrence n'est pas typé. La perte d'information est donc considérable en comparaison de l'approche syntaxique. Une solution qui permet de caractériser plus précisément le contexte d'apparition d'un mot est d'enregistrer la position des mots de son contexte au sein de la fenêtre (Rapp, 1999; Kolb, 2008). Cette manipulation permet de caractériser plus précisément le lien qui unit un mot cible et ses cooccurrents. Elle ajoute des éléments de structure – non linguistique – à une approche qualifiée de *sac de mots*. Kolb (2008) explique en partie les performances de son modèle DISCO par la prise en compte de la position des mots dans la fenêtre de cooccurrence. Gamallo Otero (2008) montre également que son système d'extraction de co-hyponymes obtient de meilleurs résultats losrque l'ordre des mots est enregistré. Ce cas de figure est illustré dans le tableau 2.2.

2.1.2.3 Comparatif

Il est difficile de comparer l'approche syntaxique et l'approche à fenêtres de mots. En effet, si elles s'appuient sur la même hypothèse de départ – les mots qui partagent les mêmes contextes ont des chances de partager des éléments de sens –, elles diffèrent sur la définition du contexte qu'elles

[6]On note $x+y$ la taille de la fenêtre de mots qui entoure un mot cible, x et y renvoyant respectivement au nombre de mots à sa gauche et à sa droite.

[7]Pour un état de la question, cf. Sahlgren (2006).

Modèle à fenêtre de mots (fenêtre de 3+3 mots)	
Cooccurrence simple	<*paysan,* COOC*, couvent*>, <*détruire,* COOC*, couvent*>, <*le,* COOC*, couvent*>, <*de,* COOC*, couvent*>, <*carmélite,* COOC*, couvent*>, <*et,* COOC*, couvent*>
+ filtrage des mots *pleins*	<*paysan,* COOC*, couvent*>, <*détruire,* COOC*, couvent*>, <*carmélite,* COOC*, couvent*>, <*piller,* COOC*, couvent*>, <*monastère,* COOC*, couvent*>
+ facteur positionnel	<*paysan, -2, couvent*>, <*détruire, -1, couvent*> <*carmélite, +1, couvent*>, <*piller, +2, couvent*> <*monastère, +3, couvent*>

TAB. 2.2 – Cooccurrents du nom *couvent* extraits de la phrase (1) en fonction de la conception du contexte adoptée.

adoptent. Il en découle des méthodes de mise en œuvre à complexité variable et des résultats de nature différente.

Du point de vue de leur mise en œuvre, l'approche à fenêtres de mots est plus facile à implémenter et plus rapide à exécuter que l'approche syntaxique, qui nécessite au préalable d'analyser syntaxiquement le corpus (qui, dans certains cas, peut être très volumineux). De la même façon, il est possible d'appliquer l'approche à fenêtres de mots sur n'importe quel type de texte pour n'importe quel type de langue. La méthode syntaxique, en revanche, est entièrement dépendante de l'analyseur utilisé. Elle est, de fait, moins facilement transposable et sujette aux erreurs d'analyse.

Grefenstette (1996) est le premier à mesurer les apports du modèle syntaxique à l'extraction de relations sémantiques. En comparant des ressources de référence – le Roget's Thesaurus, le Macquarie et le Webster's 7^{th} – à deux bases distributionnelles générées à l'aide de modèles syntaxiques et à fenêtres de mots, il montre que les performances des modèles varient en fonction de la fréquence des mots-cibles : la similarité est mieux captée à l'aide de la méthode syntaxique, mais seulement pour les mots qui ont une fréquence élevée dans le corpus. La raison en est que la méthode syntaxique ne permet d'extraire qu'un nombre réduit de contextes – le nombre de relations syntaxiques dans lesquelles entre un mot cible dans une phrase est forcément limité. Les méthodes à fenêtres de mots, en revanche, ne posent aucune limite à la taille du contexte, qui peut aller jusqu'à englober le document dans son ensemble (Salton *et al.*, 1975). De ce fait, ces méthodes se montrent plus efficaces en ce qui concerne les mots rares, pour lesquels la méthode syntaxique n'extrait qu'un nombre trop réduit de contextes. Ce problème bien connu est celui de la sparsité (déjà évoqué à la section 2.1.1). Il s'observe aussi bien sur le

corpus de seulement 3,9 megaoctets qu'a utilisé Grefenstette (1996) que sur des corpus de plusieurs dizaines de millions de mots (Weeds et Weir, 2005; van der Plas, 2008; Gamallo Otero, 2008). Van der Plas (2009) réussit à atténuer ce phénomène en mettant en place une méthode qui s'appuie sur la transitivité des relations de deuxième ordre (si A et B ne partagent pas – ou très peu – de contextes mais ont ceux de C en commun, alors il est possible que l'absence de lien AB soit due à un problème de sparsité).

D'autres études ont emboîté le pas et ont confronté les performances des deux modèles sur différentes tâches. Les résultats donnent globalement un léger avantage au modèle syntaxique. Pekar (2004) mène une évaluation d'une série de paramètres relatifs à l'ADA dans le cadre d'une tâche de classification. Ses résultats plaident en faveur de la prise en compte des relations de dépendance : le modèle syntaxique permet en effet d'obtenir des clusters de meilleure qualité qu'avec le modèle à fenêtres de mots.

Padó et Lapata (2007) montrent que la prise en compte du contexte syntaxique permet d'obtenir de meilleures performances sur des tâches d'association de mots (*semantic priming*), de détection de synonymes (questions du TOEFL) et de désambiguïsation.

Rothenhäusler et Schütze (2009) comparent les deux modèles sur une tâche de classification. Leurs résultats montrent que l'approche syntaxique permet d'obtenir des clusters d'une *pureté* supérieure malgré des contextes en quantité moindre.

Baroni et Lenci (2010) évaluent une série de modèles face à une ressource de référence comme les couples de Rubenstein et Goodenough (1965) ou sur plusieurs tâches comme la détection de synonymes (TOEFL), la classification de noms, la détection d'analogies, etc. Les résultats montrent que leur meilleur modèle, qui est syntaxique, obtient des performances comparables à des modèles à fenêtres de mots. Sans faire la démonstration d'une stricte supériorité des modèles syntaxiques sur les modèles à fenêtre de mots, Baroni et Lenci (2010) concluent donc prudemment sur une équivalence des deux approches :

> In our experiments [...], the performance of unstructured and structured models trained on the same corpus is in general comparable. It seems safe to conclude that structured models are at least not worse than unstructured models. (p. 5)

Bien que la plupart de ces études portent sur l'anglais, des travaux similaires ont été menés sur d'autres langues. Peirsman *et al.* (2007) sélectionnent un jeu de 976 mots-cibles en néerlandais pour lesquels ils extraient les 10 mots les plus distributionnellement proches en se servant d'un modèle syntaxique puis à fenêtre de mots. Une comparaison avec les relations contenues dans

la version néerlandaise d'EuroWordNet (Vossen, 1998) montre que c'est le modèle syntaxique qui permet d'extraire le plus grand nombre de relations (principalement de la synonymie). Toujours sur le néerlandais, van de Cruys (2010) obtient les mêmes résultats : il montre que le modèle syntaxique est celui qui permet d'extraire le plus grand nombre de relations recensées dans un réseau lexical néerlandais (relations de similarité) et d'obtenir les clusters de meilleure qualité. Ses performances sont à peu près équivalentes à celles du modèle à fenêtre de mots dans une tâche d'extraction de relations topicales.

Gamallo Otero (2008) montre également la supériorité du modèle syntaxique dans une tâche d'extraction de co-hyponymes en portugais.

2.1.3 Les mesures de pondération

Une fois extraits, les contextes sont pondérés. La procédure de pondération s'appuie sur l'hypothèse selon laquelle certains contextes d'un mot sont de meilleurs descripteurs de ce mot que d'autres. Par exemple, dans le cas d'une approche syntaxique, le fait que le nom *couvent* cooccurre dans le corpus Wikipédia avec *devenir*_SUJ ou *se trouver*_SUJ ne nous donne que peu d'informations sur la nature sémantique de *couvent*. Ces contextes sont très fréquents : ils apparaissent respectivement 27 488 et 16 420 fois dans le corpus et cooccurrent à au moins 5 reprises avec 1603 et 889 lemmes différents. À l'opposé, les contextes *inhumer*_À et *moine*_DE n'ont qu'une fréquence de 719 et de 749 et ne cooccurrent qu'avec un spectre beaucoup plus réduit de lemmes – respectivement 34 et 40. Leur cooccurrence avec *couvent* est beaucoup plus significative : ils sont plus caractéristiques du mot *couvent*, et donc beaucoup plus pertinents pour sa decription. Ce rapport d'exclusivité entre un mot et ses contextes peut se calculer à l'aide de mesures d'association comme le tf.idf (Spärck Jones, 1972), le t-score (Church et Hanks, 1990), le log-likelihood (Dunning, 1993) ou l'information mutuelle (Manning et Schütze, 1999). Plusieurs études – que nous ne développons pas ici – ont porté sur la comparaison des performances des différentes mesures de pondération (Curran et Moens, 2002; Evert, 2004; van der Plas et Bouma, 2004; Evert, 2008).

Au tableau 2.3, nous illustrons le fait que la fréquence absolue des cooccurrences ne permet pas de faire émerger les contextes les plus pertinents pour un mot cible donné. En effet, même si dans notre exemple, c'est *devenir*_SUJ qui cooccurre le plus avec *couvent* – 18 cooccurrences –, un contexte comme *inhumer*_À paraît plus pertinent : il apparaît à 15 reprises avec *couvent* et est moins fréquent que *devenir*_SUJ – 719 occurrences contre 27 488 –, son information mutuelle avec *couvent* est par conséquent plus élevée (6,656 contre 3,195).

		*inhumer*_À (719)	*moine*_DE (749)	*bâtir*_OBJ (2295)	*porte*_DE (5158)	*devenir*_SUJ (27 488)	*se trouver*_SUJ (16 420)
couvent (1361)	`fréq. abs.`	15	14	14	6	18	8
	`I. M.`	6,656	6,546	5,427	3,769	3,195	2,899

TAB. 2.3 – Comparaison de la cooccurrence brute et de l'information mutuelle entre le mot *couvent* et quelques-uns des ses contextes d'apparition.

Comme nous le verrons à la section 3.1.1.2, le calcul de l'information mutuelle constitue, dans de nombreux modèles, une étape indispensable de la génération de ressources distributionnelles. Ce score se calcule de la façon suivante (*freqTot* étant le nombre total d'occurrences dans le corpus et *freq(n)* la fréquence de n dans le corpus) :

$$IM(x,y) = log\left(freqTot \cdot \frac{freq(xy)}{freq(x) \cdot freq(y)}\right)$$

2.2 Mesurer la proximité distributionnelle

Après avoir extrait et pondéré les contextes d'apparition de chaque mot du corpus, l'étape suivante consiste à comparer ces contextes afin d'attribuer aux mots du corpus un score de proximité distributionnelle. Nous avons repris au tableau 2.4 les quelques contextes du mot *couvent* évoqués précédemment dans le tableau 2.3. Nous avons fait apparaître l'information mutuelle calculée entre chacun de ces contextes et les mots *abbaye*, *demeure* et *île*. On peut voir que la proportion et la valeur des contextes partagés entre *couvent* et ces trois mots est inégale :

- *abbaye* et *couvent* apparaissent dans les 6 contextes qui figurent dans l'exemple, et ce avec des scores d'information mutuelle relativement proches ;
- *demeure* n'apparaît pas dans les contextes *inhumer*_À et *moine*_DE ;
- *île* ne partage avec *couvent* que les contextes *devenir*_SUJ et *se trouver*_SUJ, qui, du fait de leur fréquence, ont un score d'information mutuelle assez faible pour ces deux mots.

Nous avons donc maintenant besoin d'un score nous indiquant que la distribution de *couvent* est plus proche de celle de *abbaye* que de celles de *demeure* et de *île*.

Nous évoquons ici dans un premier temps les deux points de vue qu'il est possible d'adopter lorsque l'on aborde la notion de proximité distribution-

	*inhumer*_À	*moine*_DE	*bâtir*_OBJ	*porte*_DE	*devenir*_SUJ	*se trouver*_SUJ
couvent	6,656	6,546	5,427	3,769	3,195	2,899
abbaye	6,648	8,042	3,677	2,462	2,58	2,907
demeure	0	0	6,685	4,937	3,957	3,98
île	0	0	0	0	2,324	2,675

TAB. 2.4 – Comparaison du score d'information mutuelle pour quelques contextes d'apparition des mots *couvent*, *abbaye*, *demeure* et *île*.

nelle (section 2.2.1). Nous faisons ensuite un point sur les différents types de mesures qui permettent de calculer cette proximité (section 2.2.2).

2.2.1 Deux conceptions de la proximité distributionnelle

Des auteurs comme Sahlgren (2006) ou van de Cruys (2010) ont montré qu'il était possible de distinguer deux types d'approche parmi les travaux qui portent sur la question de la proximité distributionnelle. En effet, même si le principe sous-jacent reste inchangé, le problème peut être abordé en adoptant un modèle soit géométrique, soit probabiliste.

2.2.1.1 Modèle géométrique

Dans le modèle géométrique, la distribution d'un mot est vue comme un vecteur dans un espace à n dimensions. Le nombre de dimensions d'un espace sémantique est égal au nombre de contextes qui permettent de caractériser un mot cible. Ce nombre est potentiellement illimité : les travaux menés en sémantique vectorielle mobilisent des espaces sémantiques dont le nombre de dimensions se compte en millions (Lund et Burgess, 1996; Landauer et Dumais, 1997; Sahlgren et Karlgren, 2005; Padó et Lapata, 2007; Baroni et Lenci, 2010; Turney et Pantel, 2010, etc.). Ces derniers seraient évidemment impossibles à représenter.

Pour expliquer ce principe, nous reprenons les illustrations de Sahlgren (2006, p. 18) en les adaptant aux exemples du tableau 2.4. Nous avons ainsi représenté à la figure 2.1 les mots *couvent*, *abbaye*, *demeure* et *île* dans un espace unidimensionnel. Tous les mots y sont situés en fonction de leur score d'information mutuelle avec le contexte *se trouver*_SUJ. Nous avons rajouté une autre dimension à la figure 2.2. La position des mots varie cette fois en fonction de leur proximité avec les contextes *se trouver*_SUJ et *bâtir*_OBJ. On peut voir que les mots *couvent* et *abbaye* sont particulièrement proches,

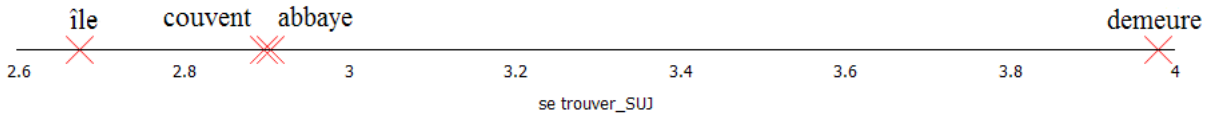


FIG. 2.1 – Positionnement des mots *couvent*, *abbaye*, *demeure* et *île* dans un espace unidimensionnel.

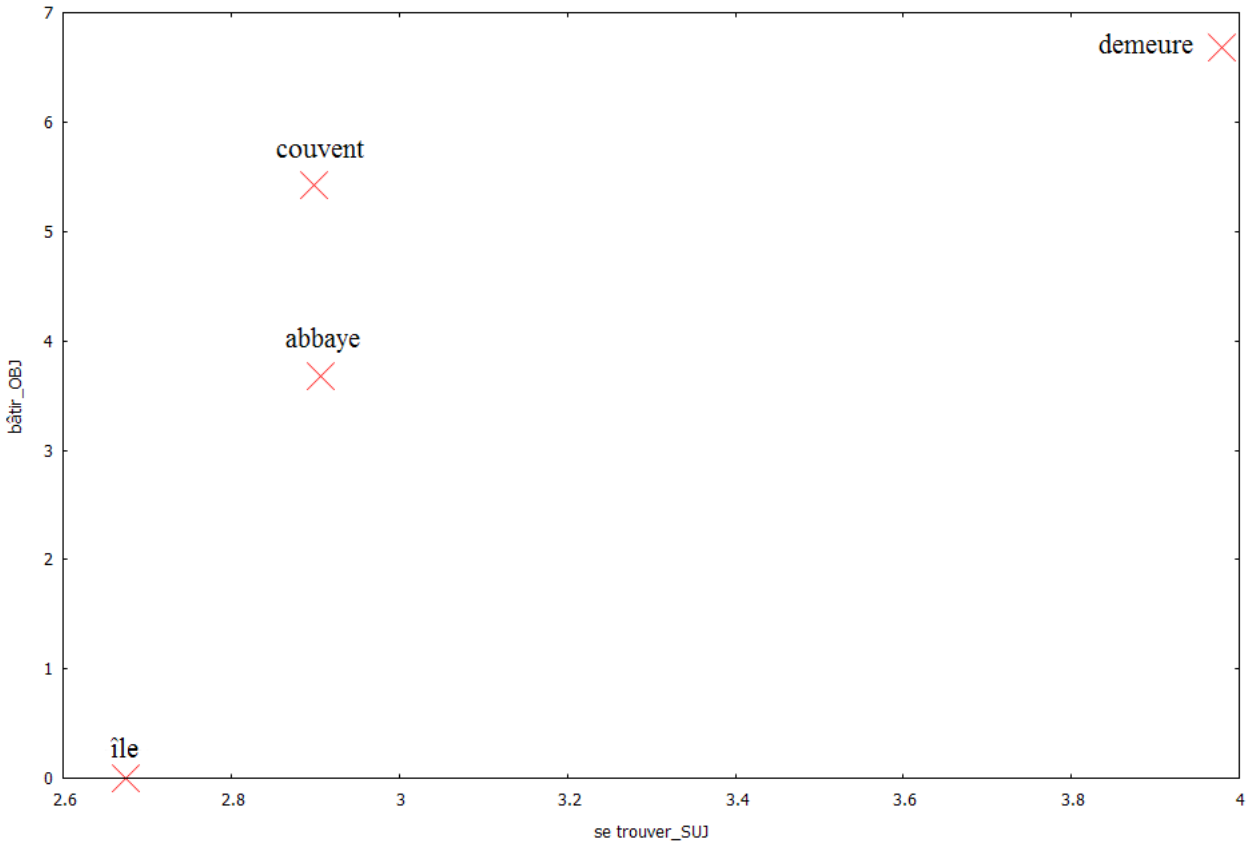


FIG. 2.2 – Positionnement des mots *couvent*, *abbaye*, *demeure* et *île* dans un espace bidimensionnel.

et ce dans les deux espaces que nous avons rapportés aux figures 2.1 et 2.2. Cette position est directement interprétable sur le plan sémantique : plus les vecteurs de deux mots sont proches dans un espace donné, plus ces mots seront sémantiquement proches. On retrouve bien ici le principe harrissien que nous avons décrit au chapitre 1. Dans le sens où c'est la position des vecteurs qui détermine le sens d'un mot, on parle de "sémantique vectorielle". Ce mode de représentation est particulièrement commode en ce qu'il permet de concevoir les contextes des mots du corpus comme des objets géométriques. De ce fait, ils sont manipulables comme tels : il est possible de calculer leur produit, leur angle, leur distance, etc. (Memmi, 2000).

2.2.1.2 Modèle probabiliste

Le modèle probabiliste peut être également utilisé pour comparer la distribution de deux mots. Il se distingue du modèle géométrique du fait qu'il se situe dans un paradigme différent : l'idée consiste ici à calculer la probabilité que deux mots aient des distributions similaires. Les distributions ne sont plus envisagées comme des vecteurs dans des espaces sémantiques mais comme des ensembles d'attributs dont il s'agit de mesurer le recouvrement (Hindle, 1990; Ruge, 1992; Grefenstette, 1994b; Lin, 1998b).

2.2.2 Les mesures de proximité distributionnelle

Plus haut dans cette section, nous avons comparé la distribution de *couvent* avec celles de *abbaye*, *demeure* et *île* dans un espace à 6 dimensions (tableau 2.4). Nous avons vu que certains mots n'apparaissaient pas dans certains contextes, et que certains contextes avaient plus de poids pour certains mots. Dans cet exemple, il était possible d'ordonner grossièrement, *à l'œil nu*, les mots *abbaye*, *demeure* et *île* en fonction de leur proximité distributionnelle avec *couvent*. Cela serait évidemment inenvisageable dans des espaces de plus grande dimension. Afin de manipuler le concept de proximité distributionnelle, il faut donc disposer d'un indice qui permet d'ordonner les couples de mots en fonction des contextes qu'ils partagent.

De nombreuses mesures ont été utilisées pour quantifier la proximité entre les mots. Nous avons vu que l'analogie géométrique permettait de mobiliser des concepts mathématiques lors de la manipulation des vecteurs contextuels. Ainsi, il est possible de calculer le cosinus de deux vecteurs afin de mesurer leur degré de similarité. Pour deux vecteurs x et y de n éléments, le cosinus

se calcule de la façon suivante :

$$cos(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}}$$

L'utilisation du cosinus en recherche d'information a été introduite par Salton et McGill (1983). Dans ses premières applications en recherche d'information, cette mesure a servi à calculer la similarité entre un vecteur constitué des mots d'une requête et les vecteurs que sont les textes d'un corpus donné. L'idée est de calculer l'angle entre le vecteur de la requête et les vecteurs de chacun des documents de la collection, puis d'ordonner ces documents par cosinus décroissant : plus ce dernier est haut, plus le contenu du texte est similaire à celui de la requête. Cette mesure a été largement utilisée (Ruge, 1995; Fung et McKeown, 1997; Landauer et Dumais, 1997, etc.) et permet encore aujourd'hui d'obtenir, sur certaines tâches, de meilleures performances que des mesures qui ont développées plus récemment (Bullinaria et Levy, 2007; Ferret, 2010; Henestroza Anguiano et Denis, 2011).

Devant le nombre de mesures qui ont été proposées par la suite[8], certaines études se sont attachées à comparer leurs performances dans des systèmes de recherche d'information (Salton et McGill, 1983), d'extraction de synonymes (notamment sur les données du TOEFL (Turney, 2001; Terra et Clarke, 2003; Ferret, 2010)) ou encore de désambiguïsation (Weeds, 2003), souvent en combinaison avec d'autres paramètres comme la méthode de pondération, afin de voir si certaines associations mesure de pondération/mesure de similarité étaient plus efficaces que d'autres (Curran et Moens, 2002; Ljubešić *et al.*, 2008; Bannour *et al.*, 2011; Panchenko et Morozova, 2012). Weeds (2003) compare également différentes mesures en fonction de jugements de similarité émis par des sujets humains ou de couples de mots recensés dans des ressources de référence. Nous aurons l'occasion de revenir plus tard sur ces modes d'évaluation.

2.2.3 La réduction de matrice

Une autre étape qui entre parfois dans la mise en place d'une analyse de corpus est la réduction de matrice. Cette opération consiste à réduire le nombre de dimensions d'un espace vectoriel. Son utilité est double : d'une part, elle permet de faciliter le traitement de la matrice – dans le cas où l'on a affaire à des données très volumineuses – et, d'autre part, de faire émerger les dimensions dont les variations sont les plus prégnantes. De ce fait, elle permet de réduire les problèmes de sparsité.

[8]Se référer à Budanitsky et Hirst (2001) ou Weeds (2003) pour un état de l'art.

L'analyse sémantique latente, ou *LSA* (Landauer et Dumais, 1997) met en œuvre une méthode de réduction de dimensions appelée *décomposition en valeurs singulières* (plus connue sous le nom de *singular value decomposition*, ou *SVD*). Dans ce cas, l'idée est de traiter une matrice mot*document afin de faire émerger les faisceaux de dimensions correspondant à des *concepts*[9]. On peut également citer le Random Indexing (Sahlgren, 2006), une autre méthode populaire de réduction de matrice qui est apparue par la suite.

Pour une explication des principes qui sont mis en œuvre, nous renvoyons le lecteur à Sahlgren (2006) ou encore à van de Cruys (2010), qui montre que la réduction de dimensionnalité n'améliore pas – voire dégrade – les performances de son système.

2.3 Constituer des classes de mots

La troisième étape dans la trilogie définie par Grefenstette (1994a) consiste à s'appuyer sur les similarités de deuxième ordre – entre les mots qui partagent les mêmes contextes d'apparition – pour mettre au jour des ensembles de mots partageant des propriétés distributionnelles. Cette démarche s'appuie sur l'hypothèse selon laquelle les classes distributionnelles ont une pertinence au niveau sémantique.

La tâche de classification est née du besoin de partitionner les masses de documents rendus disponibles par l'avènement du texte numérique. Elle consiste à *ranger* des documents dans des classes qui sont soit prédéfinies (classification supervisée), soit générées au fil de l'analyse des propriétés communes des documents (classification non supervisée, ou *clustering*). Cette approche peut également être utilisée pour mettre au jour des classes de mots (on parle alors de *word clustering*). Elle trouve notamment son utilité dans le cadre de la création d'ontologies, en ce qu'elle permet de dessiner des ensembles d'entités qui peuvent par la suite être identifiés par les experts comme des concepts du domaine.

Nous évoquons à présent quelques-unes des méthodes qui permettent de classifier des mots (section 2.3.1) avant d'aborder le problème de l'interprétation des classes générées (section 2.3.2).

2.3.1 Méthodes de classification

Nous avons vu que les méthodes de classification pouvaient être soit supervisées, soit non supervisées. Dans les deux cas, ces méthodes consistent à s'appuyer sur des calculs de proximité pour scinder un ensemble d'objets en

[9]Voir Baker (2005) pour un tutoriel.

sous-ensembles. Ici, ces objets sont des mots qui entretiennent une relation de proximité distributionnelle mesurée lors de l'étape précédente.

2.3.1.1 Classification supervisée

Dans le cas de la classification supervisée, l'idée est d'agréger des mots autour d'un ensemble de *mots amorces* (Zweigenbaum et Habert, 2006) ou *germes* (*seed words*) qui ont été annotés comme caractéristiques d'une classe ou qui possèdent des propriétés que l'on souhaite retrouver dans les mots non annotés. Habert et Zweigenbaum (2002) évoquent ces deux méthodes de classification supervisée :

- dans Riloff et Shepherd (1997) ou Roark et Charniak (1998), l'approche consiste à choisir manuellement un jeu de 5 mots amorces pour un ensemble de classes (*airplane*, *car*, *jeep*, *plane* et *truck* pour la classe VEHICLE, par exemple) et à enrichir ce jeu de façon incrémentale en l'augmentant des n mots qui ont le plus haut score de similarité avec les mots amorces. Les résultats sont ensuite soit évalués par des locuteurs (Riloff et Shepherd, 1997), soit comparés au réseau WordNet (Fellbaum, 1998), comme dans Roark et Charniak (1998) ;
- Nazarenko *et al.* (2001) sélectionnent un ensemble de termes médicaux issus de l'ontologie SNOMED et effacent leurs étiquettes. L'idée ici est d'identifier l'étiquette de ces termes en s'appuyant sur leurs proximités avec des termes étiquetés.

La classification supervisée trouve ainsi toute sa pertinence dans le cas où l'on souhaite classer les mots dans un ensemble de catégories définies *a priori*.

2.3.1.2 Classification non supervisée

Dans l'approche non supervisée, au contraire, les catégories émergent de l'analyse des données :

> Dans [le cas d'une approche non supervisée], classer signifie suivre la distribution des données dans l'espace de façon à la découper au mieux en un ensemble de groupes. (Memmi, 2000, p. 14)

On parle ici de *clusters* de mots. Les méthodes de classification non supervisées se répartissent en deux groupes : les méthodes hiérarchiques et non hiérarchiques (Memmi, 2000). Les premières permettent de produire une représentation arborescente des données – un *dendrogramme* –, dans laquelle les mots sont regroupés dans des classes imbriquées dont la granularité décroît jusqu'à arriver au niveau des mots (ou *feuilles*). Dans le cas des méthodes non hiérarchiques – méthode des k-moyennes, nuées dynamiques, centres mobiles, etc. –, les mots s'agrègent de façon itérative autour d'un ensemble prédéfini

de centres (ou *centroïdes*)[10]. L'inconvénient de cette approche est donc qu'elle suppose que l'on possède déjà une idée du nombre de classes qui vont émerger de l'analyse (il est possible de procéder dans un premier temps à une analyse hiérarchique pour avoir une estimation du nombre de clusters qui se dégagent).

Habert et Nazarenko (1996) et Fabre *et al.* (1997) choisissent de représenter les rapprochements générés par le système Zellig par des graphes dans lesquels les mots sont des nœuds et les contextes partagés des arêtes. Il est alors possible d'étudier les regroupements de mots en s'appuyant sur les concepts de *composante connexe* et de *clique* :

- une composante connexe est un sous-graphe défini par le fait que tous ses nœuds sont reliés entre eux ;
- une clique est un sous-graphe dans lequel tous les nœuds sont directement reliés à tous les autres par une arête.

Un exemple de composante connexe est donné à la figure 2.5[11]. Dans l'approche non supervisée, la démarche est plus exploratoire, elle permet de faire apparaître des classes construites en corpus. Il devient alors possible de remettre en question les classifications établies *a priori* en les confrontant avec les clusters générés par l'analyse. Le principal inconvénient de cette méthode est que les classes générées ne sont pas étiquetées, ce qui laisse une grande part d'interprétation dans l'évaluation des résultats produits. Nous abordons ce problème dans la section suivante.

2.3.2 Interpréter les classes distributionnelles

Nous décrivons ici la façon dont peuvent s'interpréter les regroupements opérés par les méthodes de classification décrites dans la section précédente. Nous évoquons ensuite les limites de l'approche à travers plusieurs études qui rapportent des divergences entre les regroupements calculés à partir du corpus et l'intuition de ce qu'un locuteur considère comme une classe sémantique.

2.3.2.1 Illustrations

Le problème de l'interprétation des classes distributionnelles ne se pose que dans le cas des approches non supervisées. Afin d'illustrer la problématique de l'évaluation des résultats fournis par les deux méthodes de classification non supervisée évoquées précédemment, nous avons appliqué ces

[10]Pour plus de détails sur la façon dont fonctionnent ces deux méthodes, se référer à Memmi (2000).

[11]Cette version de la figure est extraite de Nazarenko (2004).

méthodes à 15 des plus proches voisins du mot *couvent* dans les voisins de Wikipédia, l'une des trois ressources distributionnelles que nous utilisons par la suite. On peut voir sur le dendrogramme représenté à la figure 2.3 que les clusters générés par la classification ascendante hiérarchique[12] peuvent être interprétés de plusieurs façons. En effet, il est possible de couper les branches de cet arbre à un niveau de profondeur donné pour obtenir le nombre de classes jugé le plus pertinent. Ici, on peut par exemple élaguer l'arbre au niveau du trait en pointillés afin de diviser l'ensemble des mots en trois classes. On peut ainsi voir dans le cluster de droite que *abbaye* et *monastère* ont été rapprochés entre eux et, dans une moindre mesure, avec le mot *temple*. Dans ce cas, il est assez facile de *labéliser* le cluster, puisqu'il regroupe les noms de bâtiments religieux. Ce serait un peu moins évident pour le cluster de gauche, composé du couple *collège/lycée*, mais aussi du mot *hôpital*. Le cluster médian est encore plus hétérogène. On pourrait en effet le décomposer en 3 sous-clusters :

- le couple *fort/forteresse*, qui renvoie à des bâtiments militaires ;
- le cluster *villa/résidence/citadelle/manoir/demeure* regroupe des noms d'habitations (sauf *citadelle*) ;
- *usine*, qui se distingue de l'ensemble des autres mots du cluster.

On voit que, dans le cas de ce cluster, il aurait été plus pertinent d'élaguer l'arbre à un niveau moins élevé : en l'état, il serait en effet difficile de le labéliser en s'appuyant sur les propriétés des éléments qu'il contient. Autrement dit, ce cluster ne semble pas constituer ce que l'on définirait intuitivement comme une *classe*.

Ce constat émerge également de l'analyse de la classification des mêmes 15 voisins de *couvent* générée par une méthode non hiérarchique, celle des k-moyennes. La figure 2.4[13] montre en effet que les clusters obtenus – au vu des résultats fournis par l'analyse composante hiérarchique, nous avons choisi d'en générer 3 – sont encore moins facilement interprétables que dans le cas du dendrogramme : on retrouve la classe *collège/lycée/hôpital*, *temple* est cette fois exclu de tout cluster, et les 11 mots restants sont tous regroupés au sein d'une seule et même classe.

[12]La méthode qui a été utilisée est celle qui est implémentée dans R (R Development Core Team, 2011), c'est-à-dire la méthode de Ward. La mesure de similarité est la distance euclidienne.

[13]Ici aussi c'est l'algorithme des k-moyennes implémenté dans R qui a été utilisé.

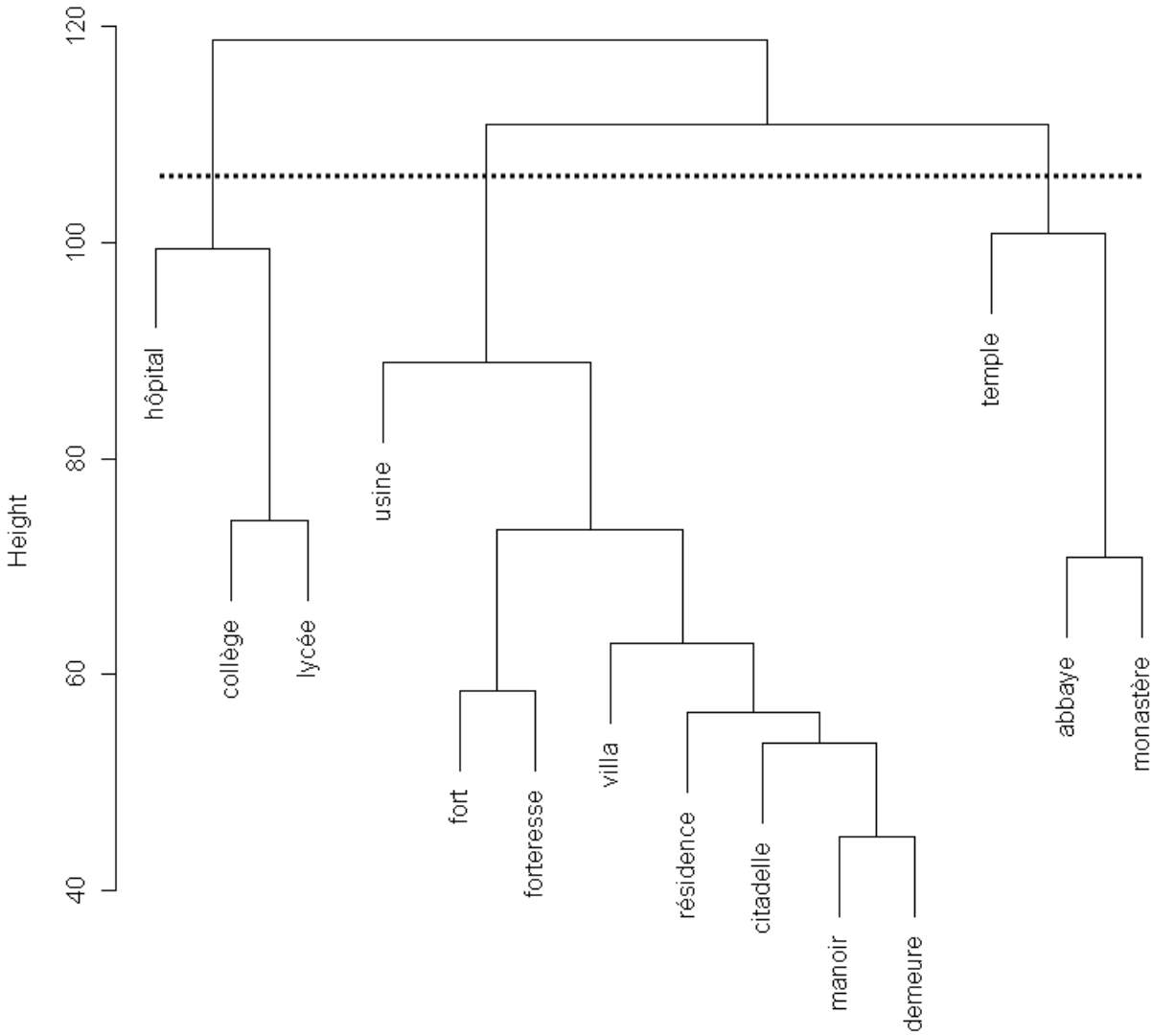


FIG. 2.3 – Classification ascendante hiérarchique de 15 voisins de *couvent*.

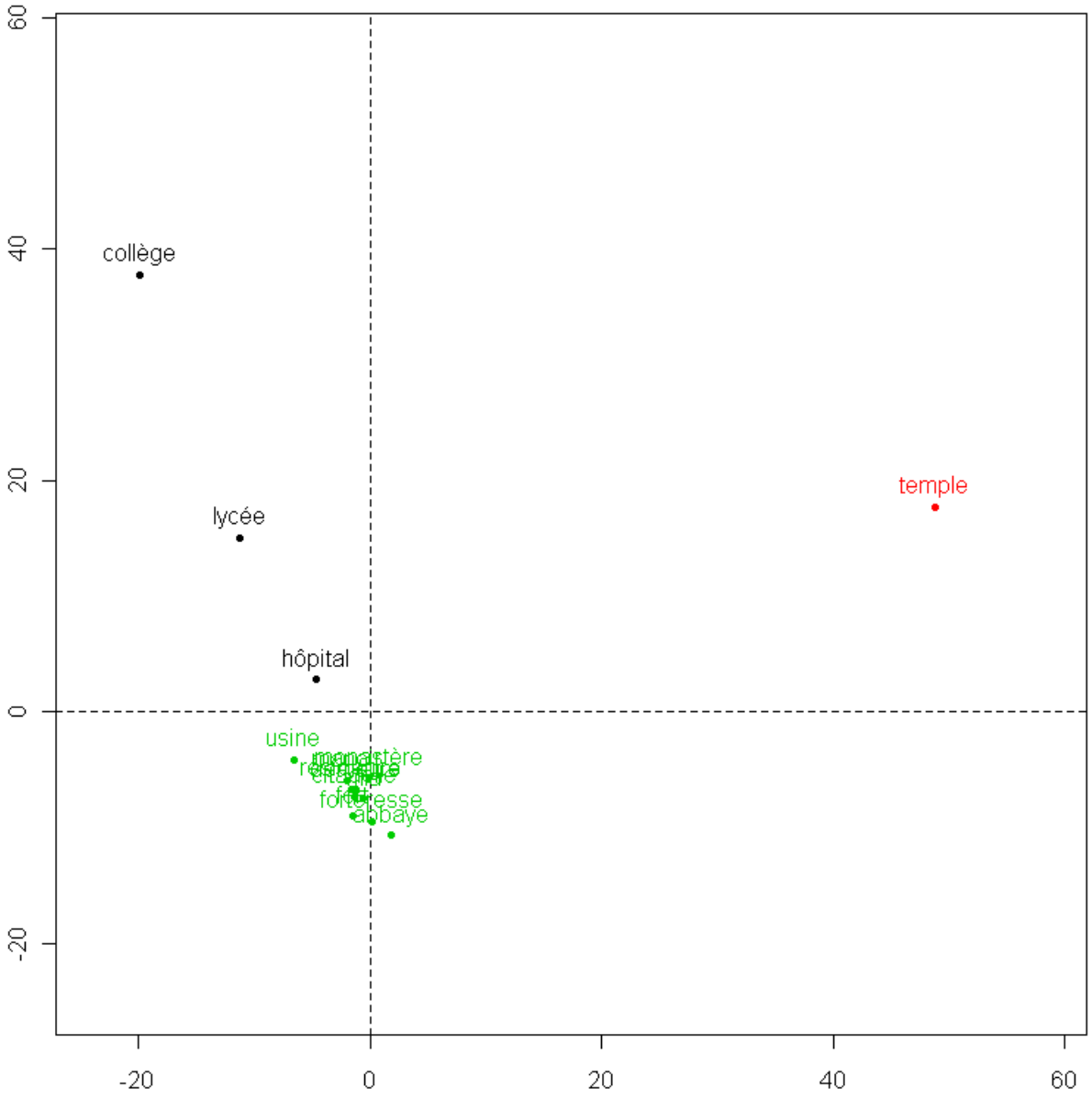


FIG. 2.4 – Classification de 15 voisins de *couvent* par la méthode des k-moyennes.

2.3.2.2 Une inadéquation entre classe distributionnelle et classe sémantique

Les travaux qui ont été menés par Harris et son équipe sur les corpus de textes médicaux reposent sur l'hypothèse d'une corrélation entre les classes distributionnelles et les classes d'objets :

> Dans un sous-langage, les classes lexicales ont des frontières relativement tranchées qui reflètent la division des objets du monde en catégories qui sont clairement différenciées dans le domaine. (Sager (1986), cité et traduit par Habert (1998))

Or, les premiers travaux dans le domaine de l'extraction automatique de classes montrent que cette hypothèse est à nuancer :

> However, no coherent and interpretable semantic classes can be built on purely endogeneous grounds. Some of the classes obtained by Bensch & Savitch (1995) or McMahon & Smith (1996) seem to be semantically sound at first sight but one must adjust their boundaries and check their consistency to turn them into actual categories. (Nazarenko *et al.*, 2001, p. 330)

Habert et Nazarenko (1996, p. 10) ont pu faire le même constat dans leur étude pionnière sur des textes en français :

> De la même manière, la syntaxe ne permet pas directement de délimiter des classes de mots reflétant une notion. Les proximités contextuelles rapprochent effectivement des mots comme les noms d'AFFECTIONS (*sténose*, *lésion*, *calcification*, *obstruction*...) qui se localisent sur des artères (*coronarien*, *circonfexe* [sic], *aortique*). Mais aucune classification ne ressort directement des graphes des figures précédentes. [...] On peut repérer des zones plus denses en liens, [...] mais les limites de ces zones restent floues.

De fait, le caractère imparfait des regroupements opérés montre qu'il est trop ambitieux de vouloir générer automatiquement des classes d'objets : les résultats obtenus sont en effet loin de constituer un matériau fini. Ils peuvent cependant être exploités comme une base de départ par les experts, pour la construction d'une ontologie, par exemple, dans une approche semi-automatique (Poibeau *et al.*, 2002). Les résultats de cette étude mettent ainsi l'accent sur le fait que la mise au jour de classes pertinentes relève avant tout d'un travail interprétatif. Pour faciliter ce travail, les auteurs choisissent de faire apparaître les contextes partagés dans les graphes produits suite aux analyses de Zellig. Comme on peut le voir dans le réseau des adjectifs rapporté à la figure 2.5, ce mode de représentation offre à l'expert une lecture claire des rapprochements effectués, ce qui lui permet de juger de leur pertinence.

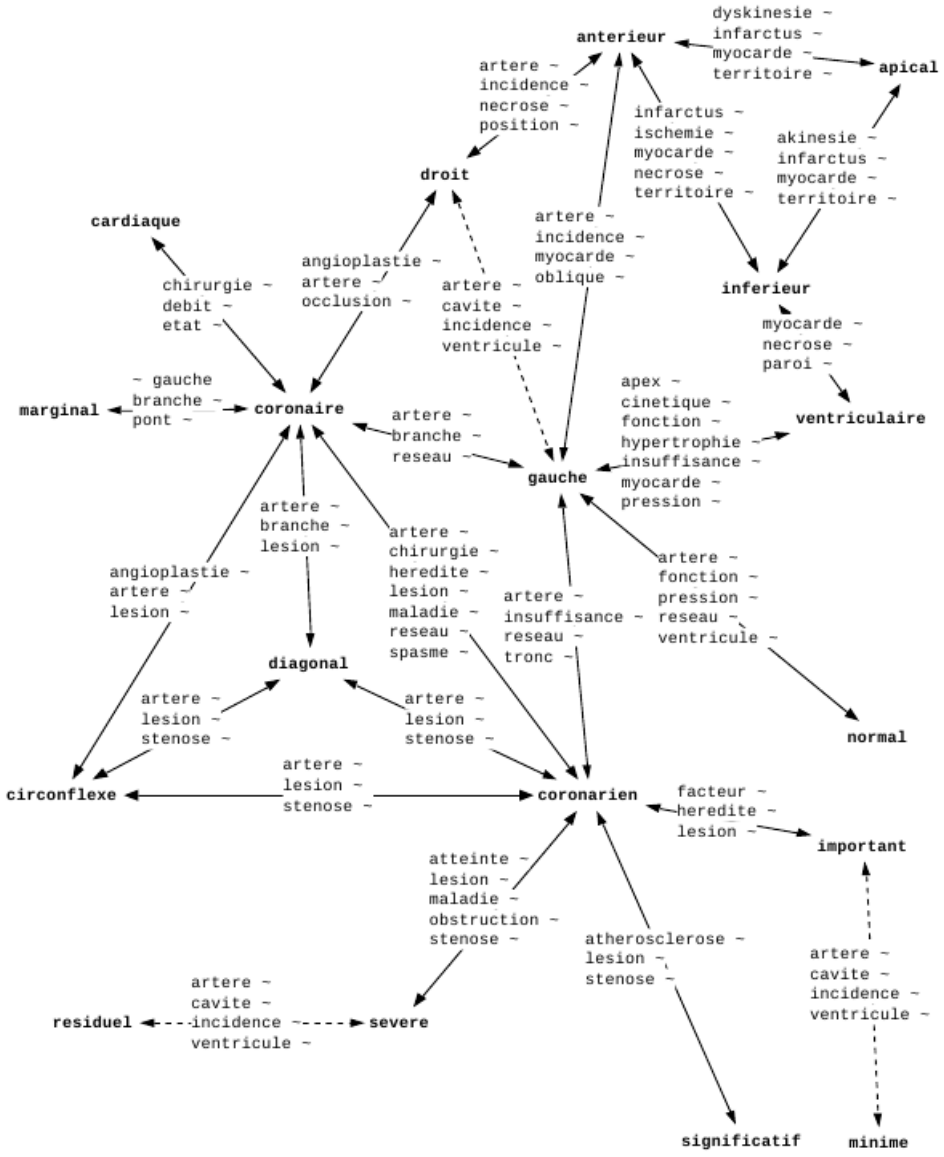


FIG. 2.5 – Extrait du graphe des adjectifs généré par Zellig à partir du corpus Menelas.

Dans le même ordre d'idée, Zweigenbaum et Habert (2006, p. 98) remettent en question le fait même d'appeler *classes* les clusters qui peuvent être mis au jour par des approches statistiques :

> C'est en outre par commodité et avec optimisme qu'on dénomme *classes* les regroupements résultants. S'ils comprennent effectivement des mots en relation de synonymie, d'hyper/hyponymie, d'antonymie, ils incluent également des relations plus complexes (méronymie ou relation de partie à tout) et des rapprochements plus douteux. Il reste donc toujours à les trier (éliminer les intrus au sein d'un regroupement et enlever les groupes "poubelles") et ensuite à les interpréter, c'est-à-dire minimalement à leur attribuer une étiquette qui "résume" leur contenu.

Ces groupes "poubelles" sont évoqués de façon différente chez Mondary (2011, p. 11) :

> Dans le cadre de la [construction d'ontologies à partir de textes] ces classes reflètent parfois des concepts que l'expert doit étiqueter, mais pas toujours : il peut arriver que les classes de mots comportent des intrus ou soient issues d'artefacts linguistiques.

Ainsi, si de nombreuses études pointent du doigt les limites de la méthode distributionnelle – et des méthodes non supervisées en général[14] (Memmi, 2000, p. 13) – la question se pose trop rarement de savoir quelle est la nature des "artefacts linguistiques" qui font que cette analyse peut faire émerger des groupes tellement hétérogènes qu'ils n'ont pas – ou peu – de pertinence du point de vue sémantique.

Cette question se pose avec d'autant plus de force à mesure que l'on s'éloigne du cadre des textes de spécialité. En effet, Fabre *et al.* (1997) et Fabre et Habert (1998) montrent que les classes que Zellig permet de mettre au jour à partir de l'analyse d'un corpus de discours de F. Mitterrand sont encore moins identifiables que celles qui ont été extraites à partir du corpus Menelas, qui relève d'une langue de spécialité. Cela est dû aux phénomènes de polysémie qui se manifestent de façon massive dans le corpus Mitterrand : un mot comme *politique* possède plusieurs facettes qu'il est possible de distinguer par l'étude des cliques qui émergent de l'analyse. Ce type de fonctionnement est exclu du corpus Menelas, dans lequel les mots sont monosémiques.

La démarche d'utiliser l'ADA pour constituer automatiquement des classes conceptuelles apparaît donc difficilement compatible avec des corpus de langue *générale*. L'hétérogénéité de la distribution des mots dans des textes comme

[14]Comme le font remarquer Habert et Zweigenbaum (2002, p. 212), la méthode distributionnelle peut en effet être considérée comme un algorithme d'apprentissage non supervisé.

ceux qui figurent dans le corpus Mitterrand appelle une approche d'un autre type, dans laquelle la dimension interprétative a une place encore plus importante. L'ADA se prêterait alors davantage à des études qualitatives basées sur la comparaison des contextes d'apparition des mots du corpus.

Dans le chapitre suivant, nous donnons un aperçu des modèles distributionnels qui ont été étudiés dans la littérature. Ces derniers varient en fonction des réglages qui ont été adoptés lors de leur mise en œuvre, lesquels ont été décrits tout au long de ce chapitre. Beaucoup d'entre eux utilisent des corpus relevant de la langue générale comme des archives de journaux, le British National Corpus (BNC) ou des textes issus du Web. Toutefois, l'approche qui est adoptée dans ces travaux ne relève pas forcément d'une démarche interprétative. En effet, les rapprochements fournis par l'analyse sont le plus souvent comparés aux couples de mots recensés dans des ressources de référence sans que ne soient décrites les raisons qui expliquent les performances du modèle mis en place.

En ce qui nous concerne, la comparaison à des ressources de référence ne constitue pas une fin en soi mais la première étape d'une série de protocoles visant à isoler des couples de mots possédant les caractéristiques que l'on cherche à observer. Comme nous le verrons au chapitre 3, l'originalité de notre travail ne réside pas dans la nature des ressources que nous étudions ni dans les paramètres qui ont été définis pour les générer. Elle se situe dans notre conception de l'évaluation – au sens large – des ressources distributionnelles, que nous envisageons comme intrinsèquement liée aux questionnements qui se posent en sémantique lexicale.

Chapitre 3

Modèles existants

Sommaire

Nous avons vu dans le chapitre précédent qu'il était possible de configurer un certain nombre de paramètres dans la mise en œuvre d'une analyse distributionnelle automatique (ADA) de corpus. La multiplication des paramètres sur lesquels il est possible d'influer constitue aussi bien un avantage qu'un des inconvénients de la méthode. Un avantage car il est possible de régler les modèles en fonction du besoin, et un inconvénient car il devient difficile de juger de l'influence d'un paramètre donné sur les résultats produits.

Nous montrons dans cette section comment ces paramètres ont été manipulés dans la littérature (section 3.1), puis lors de la génération des voisins distributionnels (section 3.2).

3.1 Panorama

Nous avons décrit, au chapitre précédent, les différents paramètres sur lesquels il est possible d'influer lorsque l'on procède à l'analyse distributionnelle automatique d'un corpus. Nous présentons ici, dans un premier temps, une série de travaux parmi ceux qui ont eu le plus d'influence dans le domaine. Nous décrivons ensuite les paramètres utilisés pour la génération des bases distributionnelles dont nous cherchons à caractériser le contenu dans la partie II (les *voisins distributionnels*).

3.1.1 Des paramètres variés pour une variété de modèles

Nous avons synthétisé, au tableau 3.1, les protocoles qui ont été mis en place dans 21 études portant sur l'ADA publiées entre 1990 à 2011. Devant la masse des travaux qui ont été menés dans ce domaine, nous avons choisi de nous focaliser sur ceux qui utilisent des modèles basés sur une conception syntaxique du contexte (afin de pouvoir comparer avec les voisins, qui s'inscrivent dans cette lignée). Nous avons également fait en sorte de mentionner en priorité les études qui sont le plus fréquemment citées dans la littérature. Ce panorama – non exhaustif – nous permet d'avoir un point de vue global sur les méthodes utilisées. On peut en effet observer plusieurs tendances dans le paramétrage des modèles.

3.1.1.1 Le type de dépendances

Même si ces travaux ont en commun une conception syntaxique des contextes, on peut les distinguer en fonction du type de dépendances retenues. On peut ainsi voir, à la colonne *contextes*, que dans les études les plus anciennes (Hindle, 1990; Pereira *et al.*, 1993; Grefenstette, 1994b), seules une ou deux relations sont prises en compte. Cela est dû au fait que l'analyse syntaxique n'en était encore qu'à ses débuts. Avec le développement des analyseurs, il est devenu possible de prendre en compte un spectre plus large de dépendances. Nous avons vu à la section 2.1.2.3 que des travaux comme ceux de Lin (1998a), Padó et Lapata (2007) ou van der Plas (2008) ont montré que la prise en compte de plusieurs relations syntaxiques augmentait la qualité des rapprochements générés. Les travaux récents s'inscrivent dans cette démarche. Toutefois, à notre connaissance, seule van der Plas (2008) s'est posée la question de savoir quelle était la conséquence de la prise en compte de tel ou tel type de dépendance sur la nature des relations lexicales captées.

étude	extraction des contextes			mesure de similarité	corpus			méthode d'évaluation
	analyseur utilisé	contextes	pondération		genre	langue	taille	
Hindle (1990)	Fidditch	Sujet, objet	IM	Hindle	Journalistique (Associated Press)	en	6 M	Illustrations
Ruge (1992)	Ruge	Modifieur	Ruge	Cosinus	Brevets	en	200 K	Jugements humains
Pereira *et al.* (1993)	Fidditch	Objet	Log-likelihood	Divergence de Kullback-Leibler	Journalistique (Associated Press)	en	44 M	Classification, prédiction d'arguments
Grefenstette (1994b)	Sextant	Sujet, objet, modifieur, complément du nom	Jaccard pondéré		Articles scientifiques, encyclopédiques, corpus Brown... (20 corpus au total)	en	De 5,5 K à 1,1 M	Tests d'association, synonymes artificiels, ressources de référence (Roget, Macquarie)
Habert et Nazarenko (1996)	Lexter	Modifieur, complément du nom	/	Seuil de x contextes partagés	Domaine médical (Menelas)	fr	20,8 K	Illustrations
Dagan *et al.* (1997)	Pattern matching	Objet	"Katz's back-off", maximum likelihood	Divergence de Kullback-Leibler, norme L1, *confusion probability*	Journalistique (Associated Press)	en	44 M	Désambiguisation
Faure et Nédellec (1998)	Sylex	Sujet, objet, modifieurs du verbe	/	Mesure dérivée de la distance de Hamming	Recettes de cuisine	fr	90 000 clauses	Illustrations
Habert (1998)	Lexter	Modifieur, complément du nom	/	Seuil de x contextes partagés	Transcriptions de discours politiques (F. Mitterrand)	fr	300 K	Illustrations
Lin (1998a)	Principar	Toutes les relations extraites	IM	Hindle, cosinus, Dice, Jaccard	Journalistique	en	64 M	Ressource de référence (WordNet et Roget)
Lee (1999)	Pattern matching	Objet	Lee	Jaccard, divergence Jensen-Shannon, norme L1, norme L2, confusion probability, τ de Kendall, *skew divergence*	Journalistique (Associated Press)	en	44 M	Désambiguisation
Bourigault (2002)	Syntex	Sujet, objet, modifieur, complément du nom	"Coefficient prox"	Jaccard	Code civil, articles scientifiques...	fr	De 145 K à 210 K	Illustrations
Curran et Moens (2002)	Sextant, Minipar	Toutes les relations extraites	T-test	Jaccard	Journalistique (Reuters), BNC	en	300 M	Ressource de référence (WordNet, Macquarie et Moby thesaurus)
Pantel et Lin (2002)	Minipar	Toutes les relations extraites	PMI	Cosinus	Journalistique (collection TREC)	en	144 M	Ressource de référence (WordNet)
Weeds et Weir (2005)	RASP	Objet	T-test, IM, log-likelihood...	Jaccard, Dice, Hindle, Lin...	BNC	en	100 M	Ressource de référence (WordNet), désambiguisation
Padó et Lapata (2007)	Minipar	Sujet, objet, modifieur, coordination... (14 relations)	Log-likelihood	Lin, cosinus, distance euclidienne, norme L1, divergence de Kullback-Leibler, *skew divergence*	BNC	en	100 M	Tests d'association, TOEFL, désambiguisation
Peirsman *et al.* (2007)	Alpino	Sujet, objet (direct et indirect), modifieur, apposition, coordination	PMI	Cosinus	Journalistique (Twente News Corpus)	nl	500 M	Ressource de référence (EuroWordNet)
van der Plas (2008)	Alpino	Sujet, objet, modifieur, coordination, apposition et complément prépositionnel	T-test, PMI	Cosinus, Dice†	Journalistique (Twente News Corpus)	nl	500 M	Ressource de référence (EuroWordNet), tâche de question-réponse, jugements humains
Rothenhäusler et Schütze (2009)	Minipar	Sujet, objet, modifieur, complément du nom, coordination... (12 relations)	T-test, g-score	Cosinus	Web (ukWaC)	en	2 MM	Ressource de référence (WordNet)
Baroni et Lenci (2010)	MaltParser	Variable en fonction du modèle	LMI	Cosinus	Web (ukWac), encyclopédique (Wikipédia), BNC	en	2 MM, 820 M, 100 M	Variable en fonction du modèle
van de Cruys (2010)	Alpino	Sujet, objet, modifieur, apposition... (8 relations)	PMI	Lin, Wu and Palmer	Journalistique (Twente News Corpus)	nl	500 M	Ressource de référence (Cornetto)
Henestroza Anguiano et Denis (2011)	MaltParser	Toutes les relations extraites	Fréquence relative, t-test, PMI	Cosinus, Lin, Jaccard	Journalistique (L'Est républicain)	fr	125 M	Ressource de référence (EuroWordNet, Wolf)

TAB. 3.1 – Protocoles utilisés dans quelques-uns des principaux travaux menés en ADA (modèle syntaxique).

3.1.1.2 La mesure de pondération

Les mesures utilisées pour pondérer les contextes (cf. section 2.1.3) varient assez peu. On peut voir, dans la colonne *pondération*, que l'utilisation du t-test est assez répandue dans les travaux récents, de la même façon que l'information mutuelle (*IM* ; Manning *et al.*, 2008) – utilisée dès les travaux de Hindle (1990) – ainsi que ses déclinaisons comme l'information mutuelle spécifique (*pointwise mutual information* ou *PMI* ; Church et Hanks, 1990) ou l'information mutuelle locale (*local mutual information* ou *LMI* ; Evert, 2008).

3.1.1.3 La mesure de similarité

Comme nous l'avons vu à la section 2.2.2, de très nombreuses mesures de similarité ont été utilisées. Malgré tout, comme on le voit à la colonne *mesure de similarité*, c'est le cosinus – utilisé dès Salton et McGill (1983) – qui est encore massivement employé. Des mesures un peu plus *exotiques* ont été mises en œuvre, mais c'est souvent dans un but comparatif.

En effet, le calcul de la similarité entre les vecteurs de mots est une étape cruciale dans le calcul d'une ressource distributionnelle et de nombreuses études ont cherché à mesurer l'influence de tel ou tel type de mesure sur les résultats fournis (Lee, 1999; Terra et Clarke, 2003; Padó et Lapata, 2007; Turney et Pantel, 2010, etc.). Weeds (2003) montre notamment que la performance de certaines mesures a tendance à varier en fonction de la fréquence des mots qu'elles rapprochent. Par exemple, la mesure de Lin, qui a été utilisée pour générer les ressources distributionnelles que nous mobilisons par la suite, est plus efficace lorsque les deux mots dont elle mesure la similarité sont de fréquences comparables. Turney et Pantel (2010) relativisent toutefois le rôle des mesures de similarité en montrant que d'autres paramètres ont une place bien plus importante dans le calcul distributionnel.

3.1.1.4 Le corpus

Nous avons recensé, dans la colonne *corpus* quelques propriétés – le genre, la langue et la taille – des corpus qui ont été utilisés dans les travaux recensés. Le constat qu'il en ressort est que la nature des corpus utilisés varie assez peu. On voit en effet que les corpus journalistiques sont ceux qui sont, de loin, les plus employés. En cela, les travaux que nous avons recensés s'éloignent de l'approche harrissienne, qui porte originellement sur des textes spécialisés. Toutefois, on peut penser que le choix de ce type de corpus a principalement été motivé par le fait que les archives de journaux, alors que le Web commençait à peine à se démocratiser, possédaient l'avantage de la masse : le

nombre de mots qu'elles contiennent se compte alors en millions, ce qui permet de répondre à la nécessité pour l'ADA de disposer de grandes quantités de données.

Cette recherche de la quantité peut pousser les auteurs à fusionner plusieurs ressources de natures différentes (Baroni et Lenci, 2010) ou, plus récemment, à recourir à des données issues du Web (Rothenhäusler et Schütze, 2009; Baroni et Lenci, 2010). Cette approche a notamment été facilitée par le développement du projet WaCky[1] (Baroni *et al.*, 2009), qui a permis de mettre à disposition des corpus issus du Web de très grande taille – l'échelle est celle du milliard de mots – dans un éventail de langues comme l'anglais le français, l'allemand, l'italien ou le norvégien[2].

La question de la quantité se trouve toutefois rattrapée par celle de la qualité. Cette tendance à amasser le plus de données possible se fait en effet au détriment d'une réflexion pourtant nécessaire autour de la caractérisation des corpus pour le calcul distributionnel. On sait que les rapprochements produits par l'ADA sont le reflet des fonctionnements des mots en corpus. Or le principal défaut des corpus issus du Web est justement qu'ils sont impossibles à caractériser (Fletcher, 2011). L'utilisation de ce type de données pour l'ADA nuit donc à l'évaluation des ressources générées en gênant l'interprétation des résultats produits. Cela concourt à augmenter l'aspect *boîte noire* de la méthode, qui perd par conséquent de son intérêt du point de vue de l'analyse linguistique.

3.1.1.5 La méthode d'évaluation

La colonne *méthode d'évaluation* indique que parmi les modes d'évaluation utilisés, c'est la comparaison à une ressource de référence – WordNet – qui est le plus employé. Cela peut s'expliquer par le besoin de comparaison des résultats obtenus avec les travaux qui ont précédé (ce qui est plus facile quand tout le monde possède le même étalon). On remarque que la désambiguïsation est également répandue. Il s'agit souvent en fait d'une tâche de *pseudo-désambiguïsation* qui consiste à modifier la forme de la moitié des occurrences d'un mot donné afin de voir si l'analyse permet de rapprocher cette nouvelle forme avec le mot original. Nous revenons sur les différentes méthodes qu'il est possible d'adopter pour évaluer une ressource distributionnelle à la section 4.2.

[1] Web-as-Corpus kool ynitiative, `http://wacky.sslmit.unibo.it`

[2] Se référer à Fletcher (2011) pour un recensement des projets similaires.

	j=1:*own*	j=2:*use*	j=1:*own*	j=2:*use*	j=1:*own*	j=2:*use*
	k=1:*bomb*		k=2:*gun*		k=3:*book*	
i=1:*marine*	40.0	82.1	85.3	44.8	3.2	3.3
i=2:*sergeant*	16.7	69.5	73.4	51.9	8.0	10.1
i=3:*teacher*	5.2	7.0	9.3	4.7	48.4	53.6

TAB. 3.2 – Tenseur de dimension 3 extrait de Baroni et Lenci (2010).

3.1.2 Repenser la modélisation des contextes

Parmi les paramètres qui n'ont été que peu remis en question figure celui de la modélisation des contextes. En effet, le modèle <*mot 1*, *mot 2*_RELATION> (<*Pierre*, *manger*_SUJ>) est pratiquement le seul à avoir été implémenté. Or, des travaux comme ceux de Turney (2008) et Baroni et Lenci (2010) ont montré qu'il était possible d'aborder la question du repérage des relations lexicales en adoptant des modèles alternatifs.

Turney (2008) propose le modèle <*mot 1*_*mot 2*, RELATION>, qui s'appuie sur le principe de la similarité relationnelle. Ce dernier est à l'œuvre dans les analogies comme *couvent*:*bâtiment*::*voiture*:*véhicule* (*couvent* est à *bâtiment* ce que *voiture* est à *véhicule*). Or, si la relation de sens qui relie *couvent* et *bâtiment* est la même que celle qui relie *voiture* et *véhicule*, alors il y a des chances que ces deux couples apparaissent dans des contextes communs comme *X est un Y*, *X et autres Y*, etc. Il devient ainsi possible de rapprocher des couples de mots – et non plus des mots seuls – en fonction de la relation qu'ils partagent.

Baroni et Lenci (2010) se sont également distingués sur ce point. Ils constatent que les différentes tâches en acquisition de relations sémantiques sont abordées avec des bases distributionnelles différentes. Partant de cela, ils proposent d'aborder ces différentes tâches à l'aide d'une seule et même ressource : la mémoire distributionnelle. Il s'agit ici d'opérer en amont, au niveau des triplets syntaxiques, pour générer un tenseur qui contient à lui seul toutes les informations nécessaires au calcul de plusieurs types de proximités distributionnelles. Nous avons rapporté au tableau 3.2 le tenseur donné en exemple dans Baroni et Lenci (2010). Il comporte 3 dimensions. De ce fait, 3 valeurs sont nécessaires pour accéder à ses *entrées* : celles de i, j et k. On désigne x_{ijk} les entrées d'un tenseur χ. Par exemple, dans le tenseur rapporté au tableau 3.2, les coordonnées x_{313} renvoient à la valeur 48,4 du triplet <*teacher,own,book*>, x_{122} renvoie à la valeur 44,8 du triplet <*marine,use,gun*>, etc.

Pour procéder au calcul distributionnel, les triplets qu'il est ainsi possible

d'extraire du tenseur peuvent être ramenés sous la forme de couples aux formats suivants :

- <*mot 1*, *mot 2*_REL>[3] → <*marine*, *gun*_USE>
  C'est ce modèle qui est le plus couramment utilisé dans la littérature. Le contexte du *mot 1* est composé d'unités hybrides formées à partir de la fusion du *mot 2* et de la relation qui le relie au *mot 1*.
- <*mot 1*_REL, *mot 2*> → <*marine*_USE, *gun*>
  Dans ce cas, c'est le *mot 1* qui est fusionné avec la relation.
- <*mot 1*_*mot 2*, REL> → <*marine*_*gun*, USE>
  Ce modèle est similaire à celui qui a été mis en place dans Turney (2008) (cf. *supra*).
- <REL, *mot 1*_*mot 2*> → <USE, *marine*_*gun*>
  Ce modèle correspond au modèle précédent, à la différence que le focus est placé sur la relation et non sur les couples de mots.

Il devient ainsi possible d'aborder à partir d'une seule ressource tout un éventail de tâches comme la résolution d'analogies ou la classification de relations – modèle <*mot 1*_*mot 2*, REL> –, la détection de synonymes – modèle <*mot 1*, *mot 2*_REL> –, l'étude de la structure argumentale des verbes – modèle <*mot 1*_REL, *mot 2*> –, etc.

3.1.3 Au delà du mot

Nous n'avons pas évoqué dans ce panorama les travaux sur l'ADA qui portent sur des problématiques liées au phénomène de compositionnalité (Erk et Padó, 2008; Baroni et Zamparelli, 2010; Mitchell et Lapata, 2010; Clarke, 2012). La récente émergence de ce champ de recherche est due au fait que les modèles distributionnels présentent une solution séduisante à la question du sens des unités complexes. Les modèles distributionnels calculent des vecteurs sémantiques pour des mots pris isolément ou des paires de mots (pour capter les analogies (Turney, 2008)). Les études qui portent sur la *compositionnalité distributionnelle* visent à induire le sens d'unités qui vont au delà du mot – syntagme, phrase, paragraphe, etc. – en se basant sur la combinaison des vecteurs des mots qui les composent.

Elles tirent leur nécessité du fait qu'il n'existe pas de corpus assez volumineux pour calculer la distribution d'entités comme des phrases. À titre d'illustration, nous adaptons l'exemple donné par A. Lenci et R. Zamparelli dans leur introduction à l'atelier *Compositionality and Distributional Semantic Models* (ESSLLI 2010, Copenhague) en comparant le nombre de résultats

[3]Nous utilisons ici notre formalisme.

renvoyés par le moteur de recherche Google[4] pour chacune des requêtes suivantes :

- "révolutionnaire" : 12 900 000
- "brûle" : 6 110 000
- "monastère" : 5 630 000

mais

- "Le révolutionnaire brûle le monastère" : 0

On peut voir que l'on rencontre un problème de sparsité : puisqu'il n'est pas possible de s'appuyer sur un corpus pour faire émerger le sens des phrases, alors ce dernier doit être reconstruit (de la même façon qu'un locuteur interprète une séquence de mots). Mitchell et Lapata (2010) illustrent la complexité de ce problème en donnant les exemples suivants :

(1) It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.

(2) That day the office manager, who was drinking, hit the problem sales worker with the bottle, but it was not serious.

Les phrases (1) et (2) contiennent les deux mêmes séries de mots, mais la façon dont ces mots sont combinés permet ici de générer deux phrases dont les sens globaux sont complètement différents. Les approches distributionnelles *traditionnelles* trouvent donc ici leurs limites dans le fait que le sens d'une phrase ne correspond pas simplement à l'addition des sens de chacun de ses mots pris isolément (Frege, 1884).

Les travaux qui sont menés actuellement sur la question de la compositionnalité ne portent pas sur des phrases aussi complexes que les phrases (1) et (2) mais sur des syntagmes. Les approches les plus répandues sont les approches additives (Foltz *et al.*, 1998) et multiplicatives (Mitchell et Lapata, 2010), qui consistent respectivement à additionner et multiplier les vecteurs afin d'en générer de nouveaux. Mitchell et Lapata (2010) accompagnent l'addition vectorielle d'une étape de pondération. Cette dernière peut consister à donner plus de poids à la tête d'un syntagme afin, par exemple, de générer des vecteurs différents pour *dog trainer* et *trainer dog*.

Mitchell et Lapata (2010) évaluent leur système en mesurant la proximité entre des syntagmes dont le sens a été constuit automatiquement et en la comparant avec des jugements humains. Par exemple, pour les syntagmes verbaux, le score de similarité entre les vecteurs de *present problem* et *face difficulty* devra être plus élevé qu'entre les vecteurs de *shut door* et *follow road*.

[4] `http://www.google.fr`

Pour une description récente des autres approches adoptées en compositionnalité distributionnelle comme les produits tensoriels (Widdows, 2008) ou la multiplication de matrices (Baroni et Zamparelli, 2010), se référer à Clarke (2012).

3.2 La chaîne Syntex-Upéry

Dans le cadre de cette thèse, nous avons travaillé avec un ensemble de ressources distributionnelles – les *voisins distributionnels* (ou simplement *voisins*) – générées au sein du laboratoire CLLE-ERSS à l'aide de la chaîne de traitement Syntex-Upéry. Le processus de génération de ces ressources a été schématisé à la figure 3.1. Dans cette section, nous décrivons les étapes de ce processus, les ressources obtenues ainsi que les travaux qui ont été menés en lien avec la chaîne Syntex-Upéry.

3.2.1 Mise en œuvre

Nous décrivons ici les deux modules qui entrent en jeu lors de la mise en œuvre de la chaîne Syntex-Upéry.

3.2.1.1 Syntex

Développé par Didier Bourigault (2007), Syntex est un analyseur syntaxique qui a été conçu comme le successeur de Lexter (Bourigault, 1994). On peut voir, dans le tableau 3.1, que Lexter a servi dans des études comme celles de Habert et Nazarenko (1996) ou Habert (1998) pour assister la génération d'ontologies. Syntex s'en différencie du fait qu'il permet d'analyser des phrases, et non plus seulement des syntagmes nominaux, à partir d'un corpus préalablement étiqueté par TreeTagger. Il procède à une analyse en dépendance dans laquelle deux mots sont reliés par une relation syntaxique, chaque mot ayant forcément un ou plusieurs dépendants dans la phrase. Nous avons représenté à la figure 3.2 l'analyse de la phrase *Les paysans révolutionnaires détruisirent le couvent* qui pourrait être produite par Syntex. Comme on peut le voir avec *paysan révolutionnaire*, Syntex procède a une extraction des syntagmes nominaux. Toutefois, ces derniers se révèlent trop rares pour pouvoir émerger lors du calcul des voisins. Ils ne constituent donc qu'une faible proportion des voisins dans nos trois ressources (moins de 4 % en moyenne).

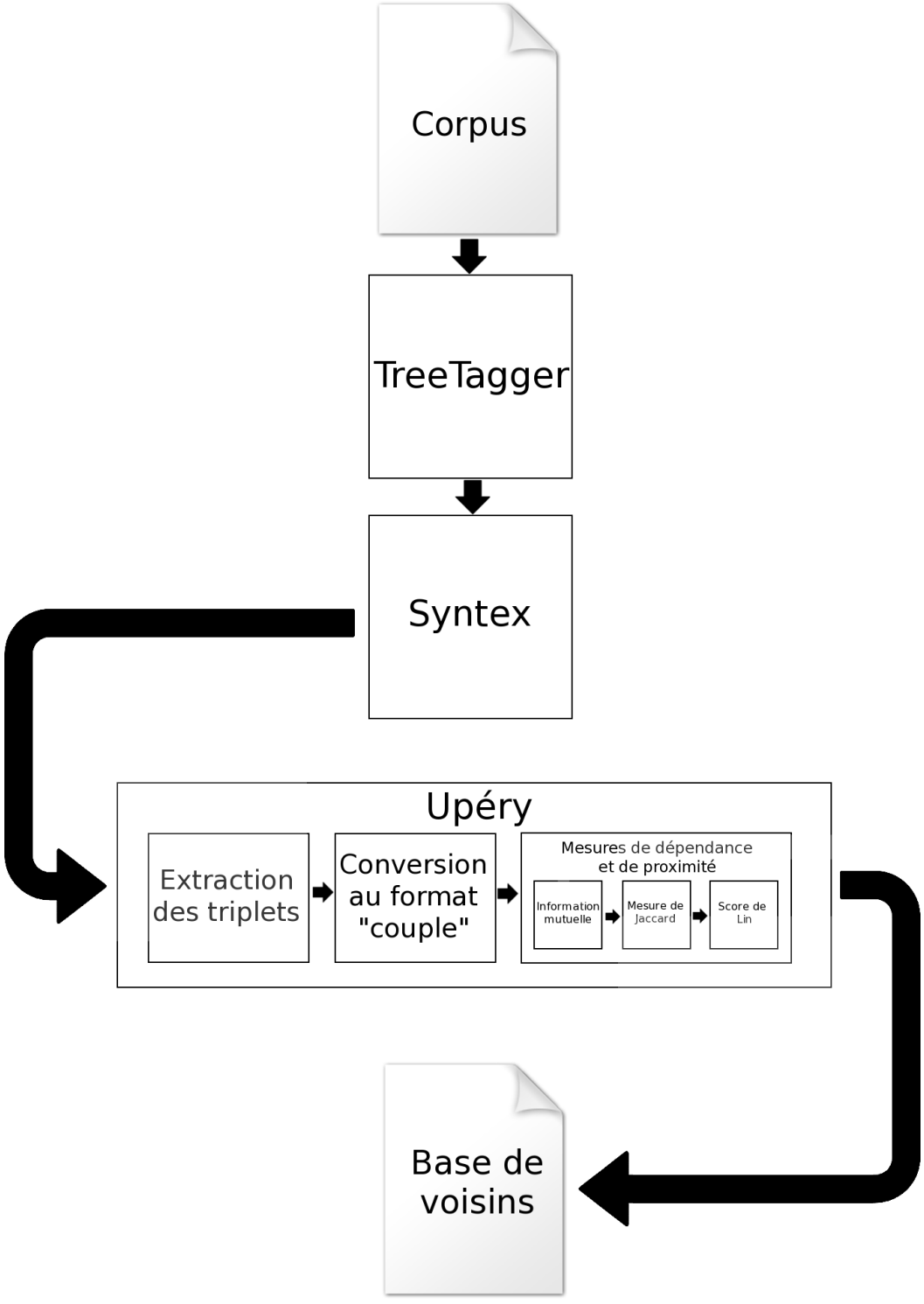


FIG. 3.1 – Étapes menant à la génération d'une base de voisins.

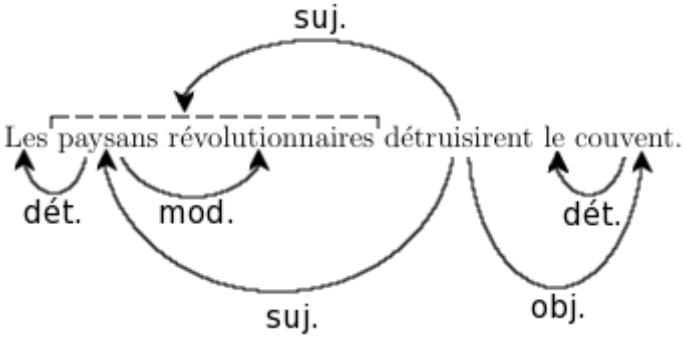


FIG. 3.2 – Exemple de phrase analysée par Syntex.

3.2.1.2 Upéry

Upéry (Bourigault, 2002) est le module qui procède au calcul des voisins. Comme son nom peut le laisser deviner, il se branche en sortie de Syntex. Le traitement des analyses générées par Syntex passe par plusieurs étapes que nous décrivons ici en nous appuyant sur la documentation rédigée par Franck Sajous (2009).

Extraction des triplets Les analyses sont dans un premier temps ramenées sous la forme de triplets syntaxiques <*mot 1*, RELATION, *mot 2*>. Seuls les triplets pour lesquels *mot 1* et *mot 2* sont des noms, des verbes, des adjectifs ou des syntagmes nominaux sont retenus. Soit les triplets suivants pour l'analyse représentée à la figure 3.2 :

- <*détruire*, SUJ, *paysan*>
- <*détruire*, SUJ, *paysan révolutionnaire*>
- <*détruire*, OBJ, *couvent*>
- <*paysan*, MOD, *révolutionnaire*>

Conversion des triplets en couples Pour les besoins du calcul distributionnel, les triplets sont d'abord réduits sous la forme de couples de type <*mot 1*, *mot 2*_REL> (cf. section 3.1.2). Dans la terminologie d'Upéry, on parle de prédicats et d'arguments :

- le prédicat résulte de la concaténation du recteur et de l'étiquette de la relation ;
- l'argument correspond au mot qui est régi.

Les résultats de la conversion des exemples précédents au format <*argument*, *prédicat*> sont donc les suivants :

- <*paysan*, *détruire*_SUJ>

– <*paysan révolutionnaire*, *détruire*_SUJ>
– <*couvent*, *détruire*_OBJ>
– <*révolutionnaire*, *paysan*_MOD>

On obtient ainsi deux types d'entités : des prédicats dont les contextes sont tous les arguments avec lesquels ils apparaissent dans le corpus, et *vice versa*. Ce formalisme va donner lieu à des rapprochements distributionnels qui opèrent à deux niveaux :

– les arguments sont rapprochés entre eux en fonction des prédicats qu'ils partagent.

couple d'arguments	prédicats partagés
paysan/ouvrier	*travailler*_SUJ, *recruter*_OBJ, etc.
couvent/temple	*bâtir*_OBJ, *piller*_OBJ, etc.

– réciproquement, les prédicats sont rapprochés entre eux en fonction des arguments partagés.

couple de prédicats	arguments partagés
*détruire*_OBJ/*construire*_OBJ	*immeuble*, *bâtiment*, etc.

Étant donné le filtrage sur les catégories grammaticales qui a été opéré en amont, on peut recenser trois types de prédicats : les prédicats verbaux, nominaux et adjectivaux. Les prédicats verbaux peuvent porter les relations suivantes :

– SUJET : <*paysan*, *détruire*_SUJ>
– OBJET : <*paysan*, *détruire*_OBJ>
– PREP : <*canon*, *détruire*_AVEC>

Dans le cas où ils sont nominaux, ils n'apparaissent que dans deux types de relations :

– MODIFIEUR : <*révolutionnaire*, *paysan*_MOD>
– PREP : <*femme*, *couvent*_DE>

Les prédicats adjectivaux ne peuvent porter qu'une seule relation :

– PREP : <*religion*, *fidèle*_À>

La relation PREP apparaît quand un complément – ou un circonstant – prépositionnel est rattaché au nom, au verbe, ou à l'adjectif. Dans ce cas, c'est la préposition qui est accolée au recteur dans le prédicat. Elle peut être soit une unité simple – *de*, *à*, *dans*, *sur*, etc. – soit une unité complexe – *lors de*, *face à*, *suite à*, *en faveur de*, etc. (on recense un total de 70 prépositions différentes dans les voisins de Wikipédia).

Ce formalisme possède l'avantage de permettre les rapprochements intercatégoriels qui font défaut aux ressources actuellement disponibles[5]. Ainsi, le prédicat verbal *prier*_OBJ se retrouve associé à des prédicats nominaux comme *culte*_DE ou *sanctuaire*_DE, avec lesquels il partage des arguments comme *dieu*, *déesse*, *vierge*, *saint*, etc.

La fréquence et la productivité des arguments et des prédicats sont ensuite calculées. Le score de productivité, que nous remobiliserons par la suite, correspond, pour un prédicat, au nombre d'arguments différents avec lesquels il apparaît dans le corpus (et réciproquement pour les arguments).

Calcul de l'IM et mesure de similarité Dans un premier temps, le programme calcule l'information mutuelle entre chaque couple prédicat/argument et argument/prédicat. Ce score n'entre pas en compte dans le calcul des voisins. Nous verrons toutefois qu'il nous permettra de mener des analyses plus fines des contextes d'apparition des mots (cf. chapitres 5 à 8). Nous ne revenons pas sur les modalités de calcul de l'IM, qui sont développées à la section 2.1.3.

Les mesures de similarité qui sont calculées sont le Jaccard et le score de Lin (1998b). Pour des raisons de clarté, nous illustrons les descriptions que nous faisons du Jaccard et du Lin en prenant seulement les prédicats pour exemple. Toutefois, ces calculs sont strictement identiques lorsqu'ils concernent les arguments.

Le calcul du Jaccard s'appuie sur le score de productivité, évoqué plus haut. Soient deux prédicats p_1 et p_2, leurs productivités respectives $prod(p_1)$ et $prod(p_2)$, et $a(p_1, p_2)$ le nombre d'arguments qu'ils ont en commun :

$$prox_{jacc}(p_1, p_2) = \frac{a(p_1, p_2)}{prod(p_1) + prod(p_2) - a(p_1, p_2)}$$

Ce score varie entre 0 et 1 (deux mots ayant des distributions strictement identiques ayant un score de 1).

Le calcul du Lin se déroule en plusieurs étapes. Une première étape consiste à calculer la *quantité d'information* (QI) de chaque prédicat, c'est-à-dire le rapport du nombre d'arguments avec lesquels il se combine dans le corpus, sur la totalité des arguments/prédicats avec lesquels il pourrait se combiner. Soit un prédicat p, $a(p)$ le nombre de ses arguments dans le corpus et $n(p)$ le nombre de ses arguments potentiels (le nombre total de noms, pour un prédicat verbal, par exemple) :

[5]EuroWordNet contient des couples intercatégoriels mais ces derniers ont été ajoutés parce qu'ils partagent des similarités morphologiques.

$$QI(p) = -log\left(\frac{a(p)}{n(p)}\right)$$

À partir de ce score, on calcule :

- $sommeQI(p)$ la somme des QI de tous les arguments que prend p ;
- $sommeQIcommun(p1, p2)$ la somme des QI de tous les arguments que partagent deux prédicats $p1$ et $p2$.

Le score de Lin entre deux prédicats peut maintenant se calculer de la façon suivante :

$$sim_{lin}(p1, p2) = \frac{2 \cdot sommeQIcommun(p1, p2)}{sommeQI(p1) + sommeQI(p2)}$$

Comme le Jaccard, ce score varie de 0 à 1.

3.2.2 Ressources générées : les voisins de *

Dans le cadre de cette thèse, nous utilisons trois ressources distributionnelles qui ont été produites par la chaîne Syntex-Upéry à partir de trois corpus de différentes natures :

- un corpus d'environ 262 millions de mots constitué de l'intégralité des articles de l'encyclopédie en ligne Wikipédia[6] dans sa version de juin 2008 ;
- un corpus comprenant l'ensemble des articles parus dans le journal Le Monde sur une période de 10 ans – de 1991 à 2000 – soit environ 200 millions de mots ;
- un corpus d'environ 30 millions de mots constitué de 515 romans datant du XX^{e} siècle issus de la base Frantext[7].

Ces trois corpus ont donc permis de générer les ressources appelées *voisins de Wikipédia* (VDW), *voisins de Le Monde* (VDLM) et *voisins de Frantext* (VDF).

La taille de ces corpus paraît modeste en comparaison de ceux qui ont pu être utilisés dans des travaux comme ceux de Turney (2008), Rothenhäusler et Schütze (2009) ou Baroni et Lenci (2010), qui utilisent les corpus issus du Web, dont le nombre de mots se compte en milliards. La raison pour laquelle nous avons malgré tout travaillé avec ces ressources est double. D'une part, elles étaient les seules disponibles au laboratoire CLLE-ERSS au début

[6] `http://fr.wikipedia.org/`

[7] `http://www.frantext.fr`

de notre thèse. Syntex n'étant plus librement utilisable au sein du laboratoire, nous n'aurions pas pu relancer le processus de calcul des voisins sur de nouvelles données et l'utilisation d'un autre analyseur aurait nécessité l'adaptation d'Upéry. D'autre part, nous nous situons dans une démarche qui s'appuie notamment sur des retours au contexte pour expliquer les phénomènes observés. Or, comme nous l'avons évoqué à la section 3.1.1.4, l'utilisation de corpus issus du Web complique ce type d'approche. Les corpus que nous utilisons ont donc l'avantage qu'ils appartiennent à des genres identifiés et connus.

Les modalités qui ont permis de calculer les trois bases de voisins varient sur trois points :

- alors que la fréquence minimale requise pour la prise en compte des triplets dans le calcul distributionnel est de 5 dans les VDW et les VDLM, elle est de 3 dans les VDF ;
- un couple de voisins doit au minimum partager 5 contextes d'apparition différents dans les VDW et les VDLM. Ce seuil est de 3 dans les VDF. Ces deux premières différences entre les VDW/VDLM et les VDF sont dues au fait que le corpus Frantext est d'une taille beaucoup plus réduite que celles des deux autres corpus ;
- un seuil de 0,1 sur le score de Lin a été appliqué aux VDW alors que les VDLM et les VDF ont été filtrés en fonction du Jaccard (respectivement à 0,1 et 0,01).

La différence de taille entre les trois corpus utilisés se répercute sur celle des trois bases de voisins. On compte ainsi :

- 3 922 657 couples dans les VDW ;
- 5 525 480 couples dans les VDLM ;
- 792 356 couples dans les VDF.

Ces trois bases sont consultables sur la plate-forme REDAC[8]. Étant donné le caractère propriétaire de l'analyseur Syntex, ces ressources ne peuvent pas être téléchargées librement. Cela pose évidemment un frein à la diffusion de la méthode distributionnelle, en plus de constituer une entrave à la reproductibilité des manipulations que nous effectuons par la suite. Comme nous l'avons indiqué plus haut, les raisons qui nous ont toutefois poussé à travailler sur ces données sont liées à un critère de disponiblité. À l'heure actuelle, le laboratoire s'est doté de plusieurs ressources distributionnelles – qui seront librement distribuables – calculées à partir des corpus Wikipédia et Frantext à l'aide de l'analyseur Talismane[9] (Urieli et Tanguy, 2013).

[8] `http://redac.univ-tlse2.fr/index.html`

[9] `http://redac.univ-tlse2.fr/applications/talismane.html`

3.2.3 Applications liées aux voisins distributionnels

L'analyseur a été conçu pour pouvoir traiter des textes appartenant à des genres variés. Ainsi, couplé au module Upéry, il a par exemple permis d'assister la génération d'ontologies exploitables dans les trois domaines suivants (ces travaux sont mis en parallèle dans Bourigault *et al.* (2004)) :

- la fabrication et l'utilisation de la fibre de verre (Aussenac-Gilles *et al.*, 2003), dans un contexte de classification automatique de documents ;
- la traumatologie en réanimation chirurgicale (Le Moigno *et al.*, 2002), avec le but d'améliorer le codage des actes médicaux ;
- le Droit (Bourigault et Lame, 2002), le but étant d'apporter à un moteur de recherche juridique de nouvelles fonctionnalités comme l'expansion de requêtes ou la catégorisation de document.

Outre l'intérêt de ce système pour la construction de ressources termino-ontologiques, l'approche adoptée dans ces travaux fait de la chaîne Syntex-Upéry un outil de choix pour l'étude de phénomènes linguistiques. Dans leur travail sur les textes juridiques, Bourigault et Lame (2002) montrent par exemple que le nom *enfant* est voisin de *mineur*, *époux* et *conjoint* dans le Code civil et de *apprenti* et *salarié* dans le Code du travail. Ces catégorisations concurrentes sont le reflet du fait que des fonctionnements différents sont à l'œuvre dans ces deux types de textes, qui appartiennent pourtant à un même sous-domaine (celui des textes juridiques).

Parmi les études qui ont utilisé la chaîne Syntex-Upéry dans une visée principalement orientée vers la recherche linguistique (e. g. qui ne répondaient pas à un besoin ponctuel d'une ontologie opérationnelle de la part d'un partenaire externe), on peut citer les travaux de Bourigault et Galy (2005) qui ont montré la faiblesse du recouvrement entre les VDLM et le dictionnaire de synonymes du CRISCO (Manguin, 2002), ou ceux de Tutin (2007) dont le but était de faire émerger des classes de mots relatives à l'activité de recherche scientifique à partir d'un corpus d'articles de linguistique, de médecine et d'économie.

Les voisins ont également servi dans plusieurs travaux orientés vers l'analyse du discours. Par exemple, dans Vergez-Couret et Adam (2012), les auteurs s'appuient sur l'idée selon laquelle deux propositions entretenant une relation d'élaboration ont tendance à partager des mots sémantiquement reliés, comme dans l'exemple ci-dessous (nous soulignons) :

> (a) United Fruit Company *investit* dans le pays, (b) en *achetant* des parts dans le chemin de fer, l'électricité et le télégraphe.

Une base de voisins – les VDW – est alors mobilisée pour repérer ce type de liens. Dans Adam et Morlane-Hondère (2009) et Adam (2012), les VDW sont utilisés pour leur potentiel à repérer des zones de cohésion lexicale. Ils

sont notamment évalués en comparaison à d'autres ressources lexicales dans un système de segmentation thématique.

Ce mode d'évaluation, dit *par la tâche* (*task-based*), ainsi que les différentes façons dont il est possible d'aborder la question de la qualité d'une ressource comme les voisins distributionnels sont abordés au chapitre suivant. Nous y comparons les trois bases de voisins évoquées plus haut à des lexiques de synonymes, d'antonymes, d'hyperonymes et de méronymes. Comme nous l'avons vu dans le panorama présenté à la section 3.1, ce mode d'évaluation est le plus répandu. Il nous permet de donner un aperçu global du contenu des bases de voisins. Ce chapitre s'insère dans la partie II : il sert de préliminaire aux chapitres 5 à 8, dans lesquels nous mettons en place des expériences visant à étudier les manifestations en corpus de chacune des relations lexicales évoquées ci-dessus.

Deuxième partie

Les voisins distributionnels comme observatoire des relations lexicales en corpus

Chapitre 4

Caractériser les voisins distributionnels

Sommaire

Dans la partie précédente, nous avons retracé la genèse de la méthode distributionnelle, de ses fondements théoriques aux mesures de proximité utilisées lors de sa mise en œuvre. Nous avons vu que tout l'intérêt de cette méthode repose sur son potentiel à faire émerger des relations de sens. Or, la nature même de ces relations n'a pas été évoquée. La raison en est que, malgré la popularité de cette méthode, la nature des relations qu'elle permet d'extraire est encore assez mal définie. Comme nous l'évoquions dans la fin du chapitre précédent, cela est dû au fait que les méthodes d'évaluation traditionnelles ont négligé le côté interprétatif des résultats au profit de méthodologies plus facilement systématisables.

Dans la section 4.1, nous donnons un aperçu de la variété des relations qu'il est possible d'identifier parmi les associations produites par la chaîne Syntex-Upéry. Nous montrons que cette variété complique la question de l'évaluation, que nous abordons à la section 4.2. À la section 4.3, nous adoptons un mode d'évaluation bien connu – la comparaison à une ressource de référence – pour donner un aperçu global du contenu des bases de voisins dont nous disposons. Nous concluons ce chapitre par une section 4.4 qui s'inscrit dans le prolongement de la démarche entamée à la section précédente : après avoir comparé les bases de voisins dans leur ensemble, nous procédons à des mesures de recouvrement plus ciblées afin de mesurer l'influence de certains critères sur la composition des voisins.

4.1 Que cherche-t-on à extraire ?

Le non-typage des rapprochements générés par l'ADA est certainement le plus gros inconvénient de cette méthode. En effet, contrairement à d'autres méthodes d'extraction de relations comme les méthodes par pattern-matching (Hearst, 1992; Morin, 1999) qui permettent de cibler le type de relation à extraire, l'ADA ne distingue les couples rapprochés qu'en fonction de leur proximité distributionnelle. Ainsi, le score de similarité indique dans quelle mesure deux mots sont proches, mais cela ne nous informe en rien sur la nature sémantique de cette proximité.

4.1.1 Similarité *vs* proximité sémantique

À la section 2.1.2, nous avons évoqué les résultats de certains travaux suggérant que le choix des paramètres liés à la nature des contextes extraits (modèle syntaxique *vs* sac de mots) pouvait avoir une influence sur la nature des relations extraites. On considère ainsi que les modèles syntaxiques sont plus adaptés pour capter des relations de similarité sémantique alors que les

modèles *sac de mots* ont tendance à extraire des relations de simple proximité.

La distinction entre les relations de *similarité* et de *proximité* sémantique est posée chez Resnik (1995), puis réexploitée dans Budanitsky et Hirst (2006). Elle s'appuie sur le principe selon lequel deux mots peuvent être reliés sémantiquement sans être similaires. Ainsi, des couples comme *moine/prêtre* ou *église/couvent* relient des entités qui sont similaires : respectivement des individus et des bâtiments (religieux). Ce n'est pas le cas dans des couples comme *moine/couvent*, *église/religion*, ou *prêtre/prier*, qui relèvent de relations associatives au sens large : les entités reliées ne sont pas de même nature (un individu et un bâtiment dans le premier cas, un bâtiment et une notion abstraite dans le deuxième, un individu et une action dans le troisième). Il est intéressant de noter que la dichotomie *similarité vs proximité* est assez proche de l'opposition entre les relations *étroites* (*tight relations*) et *lâches* (*loose relations*) évoquée chez Kilgarriff et Yallop (2000).

Halliday et Hasan (1976) ont évoqué le rôle des relations de proximité, qu'ils appellent *non systématiques*, dans la formation de la cohésion textuelle. Dans leur étude, cette relation recouvre des cas comme :

- *abeille/miel*, qui peut être identifiée comme une relation entre producteur et produit ;
- *jardin/creuser*, qui relie un nom de lieu à une action qui peut être effectuée dans ce lieu ;
- *rire/blague*, qui peut être assimilée à une relation de cause/conséquence.

Hoey (1991) montre même que ces relations participent davantage à générer de la cohésion lexicale dans les textes que les relations de synonymie, d'hyperonymie, etc. (ce qui correspond aux résultats obtenus par la suite par Morris et Hirst (2004); Morris (2007)).

Selon Budanitsky et Hirst (2006), la similarité est portée par les relations de synonymie et d'hyperonymie alors que la proximité recouvre la méronymie, l'antonymie, les relations fonctionnelles ainsi que les relations dites *non classiques*. Cette dichotomie est discutable. On peut, par exemple, s'étonner de voir une relation comme l'antonymie figurer parmi les relations de proximité alors que des auteurs comme Murphy (2003) ont montré que les antonymes partageaient la plus grande partie de leur sens, la différence ne portant que sur une seule de leurs "dimensions" sémantiques : *chaud* et *froid*, par exemple, ne se distinguent que par leur polarité. Toutefois, le problème principal de cette dichotomie est que le spectre des relations considérées comme relevant de la proximité est particulièrement flou. On a donc ici affaire à une classification qui oppose les relations d'identité que sont la synonymie et l'hyperonymie et une nébuleuse de relations non identifiées manifestant une certaine proximité sémantique entre deux mots. Il est intéressant de noter que ce cas de figure se retrouve ailleurs dans la littérature.

C'est notamment le cas de l'opposition entre les relations dites *classiques* et *non classiques*. Cette opposition, évoquée chez Budanitsky et Hirst (2006), vient des travaux de Morris et Hirst (2004) et Morris (2007), qui se sont eux-même inspirés de Lakoff (1987). Ils montrent que 62 % des relations sémantiques identifiées dans des textes par des locuteurs ne relèvent ni de la synonymie, de l'antonymie, de l'hyperonymie ou de la méronymie (relations dites *classiques*). Morris (2007) identifie 20 types de relations non classiques parmi les associations repérées par les locuteurs dans une série de textes, dont les relations de :

- lieu (*professeur/lycée*, *sans-abri/soupe populaire*) ;
- conséquence (*bouteille/saoul*, *amour/marié*) ;
- nécessité (*richesse/argent*, *revenu/travail*).

Comme le fait remarquer Murphy (2003), c'est la Théorie Sens-Texte (Mel'čuk, 1988) qui "détient le record de la théorie dont le lexique compte le plus grand nombre de relations identifiées" (p. 69, notre traduction). Comme son nom l'indique, la théorie Sens-Texte (TST) postule une primauté du sens, qui peut se manifester dans les textes sous différentes formes. De fait, le lexique se retrouve au cœur de la description linguistique. Le lexique de la TST s'appelle le Dictionnaire Explicatif et Combinatoire (Mel'čuk *et al.*, 1999). Il est "le répertoire des significatons de la langue" (Polguère, 1998, p. 17). L'aspect *combinatoire* de ce dictionnaire repose sur une soixantaine de *Fonctions Lexicales* qui permettent de décrire la façon dont se combinent les unités lexicales aussi bien sur le plan syntagmatique que paradigmatique. Par exemple, voici trois des fonctions lexicales de *navet* – dans son sens premier :

- **Syn**(*navet*)=*rutabaga*
- **Géner**(*navet*)=*légume*
- **Mult**(*navet*)=*botte* [*de* ∼s]

La première fonction exprime une relation de synonymie entre *navet* et *rutabaga*, la deuxième une relation de généricité (ou d'*hyponymie*) entre *navet* et *légume*. La troisième fonction exprime le fait que l'unité lexicale à employer pour désigner un ensemble de navets est *botte* (*de navets*). De la même façon : **Mult**(*vache*)=*troupeau*, **Mult**(*chien*)=*meute*, **Mult**(*bateau*)=*flotte*, etc. Les fonctions lexicales recensées se veulent universelles, elles ne sont pas limitées à une seule langue. De plus, le fait qu'elles peuvent être combinées entre elles permet d'éviter l'ajout constant de nouvelles fonctions, ce qui rendrait vite le modèle inutilisable.

4.1.2 Les relations *ad hoc*

Le concept de catégorie *ad hoc* a été développé par le psychologue Lawrence B. Barsalou (1983), qui montre qu'en fonction du besoin, les mots peuvent être classés dans des catégories comme, par exemple, "les choses que l'on peut vendre dans un vide-grenier", qui diffèrent des catégories préétablies comme celle des fruits ou des meubles. Morris (2007) rencontre ce cas de figure dans sa typologie à travers la relation de *co-hyponymie ad hoc*. Cette relation relie des couples comme :

- *mots croisés/livres*, qui ont été regroupés parce qu'il sont tous les deux des passe-temps qui impliquent des mots ;
- *bouteilles/colle*, qui sont des types de détritus ;
- *ticket de caisse/enveloppe*, qui sont des types de papiers que l'on jette à la poubelle.

On peut parler de relation *ad hoc* dans le sens où ces couples entretiennent des relations tellement liées au contexte qu'il serait impossible d'en faire l'inventaire : ce sont des co-hyponymes qui appartiennent à des catégories créées par le texte. Nous aurons l'occasion de revenir dans les chapitres suivants sur la question de l'influence du contexte – ou du corpus – sur la catégorisation des mots.

Dans le cadre d'une étude portant sur la fonction discursive des antonymes en contexte, Jones (2002) et Jones *et al.* (2012) extraient d'une série de corpus (journalistique, romans, Web, CHILDES, BNC, etc.) les phrases dans lesquelles apparaissent des couples d'antonymes dit canoniques (dont l'opposition est immédiatement reconnaissable par les locuteurs). Ils identifient plusieurs fonctions, parmi lesquelles la fonction *auxiliaire* (*ancillary antonymy*). Ce cas de figure se définit par le fait que, dans une phrase, une paire A constituée d'antonymes dits *canoniques*[1] engendre – ou renforce – l'interprétation antonymique de deux mots d'une paire B. Dans les exemples suivants, la paire A apparaît en gras, la paire B en italique :

(1) As the old adage put it, *oppositions* do not **win** elections ; *government* **lose** them.

(2) The *teacher* is **active** and the *student* **passive**.

(3) *Milk* is **good** for you but *gum* is **bad** for you.

(4) The new edition appeared in the United States about two weeks ago ; when I heard the news of the coup it seemed **bad** news for *democracy*, but very **good** news for the *book*.

[1] Nous aurons l'occasion de revenir sur ce concept dans le chapitre 6.

	classique	non classique
en langue	*moine/prêtre*, *église/couvent*	*rire/blague*, *professeur/lycée*
en contexte	*ticket de caisse/enveloppe*, *livre/démocratie*	*revenus/enquête*, *psychologues/intérêt*

TAB. 4.1 – Classification de quelques exemples de couples.

On peut considérer qu'on a affaire à une relation *ad hoc* dans le sens où c'est l'opposition entre les mots de la paire A qui sert de support à l'identification d'un lien de sens entre les mots de la paire B. Dans le cas où les mots de la paire B présentent un contraste *a priori*, la paire A vient renforcer cette opposition (exemples (1) et (2)). Dans d'autres cas, l'interprétation contrastive est générée de toute pièce par la paire A (exemples (3) et (4)). Ce type de constructions est particulièrement fréquent : la fonction auxiliaire est celle qui se manifeste le plus fréquemment dans les phrases extraites par Jones (2002).

On peut également citer les travaux de Adam (2012), qui montre que le fait de projeter sur un corpus des couples de voisins distributionnels permet de se faire une idée de la raison pour laquelle ces couples ont été générés. Par exemple, les mots *insecte* et *racine*, identifiés comme des voisins distributionnels, ne semble pas entretenir de relation de sens *a priori*. Or, le contexte ci-dessous – issu du corpus Wikipédia – nous permet de voir que ces deux mots peuvent être interprétés comme des co-hyponymes de la catégorie des aliments du hérisson, qui est générée par le texte :

> Bien que faisant partie des insectivores, les hérissons sont quasiment omnivores. Ils se nourrissent d'**insectes**, [...] de **racines**, de melons et de courges.

Nous avons ici vu plusieurs cas où le contexte permettait de créer une relation de sens entre deux mots qui n'en entretenaient pas *a priori*. Le fait qu'une relation soit contextuelle ou soit établie en langue est indépendant de la nature de la relation. Afin d'illustrer ce principe, nous avons classé dans le tableau 4.1 quelques-uns des exemples qui ont été cités précédemment en fonction de leur nature – classique *vs* non classique – et selon qu'elles sont établies en langue ou ont été générées en contexte. Les deux exemples de relations non classiques contextuelles qui figurent dans le tableau n'ont pas été commentés jusque-là. Ils sont extraits de Morris (2007). Les locuteurs

qui ont rapproché ces couples ont justifié leur démarche de la façon suivante (notre traduction) :

- *revenus/enquête* : “Les revenus sont parfois utilisés comme un sujet d'étude à l'aide d'enquêtes” ;
- *psychologues/intérêt* : “Les psychologues portent un intérêt à l'étude de la psychologie et des gens”.

4.1.3 Des voisins hétérogènes

Comme nous l'avons vu plus haut, le non-typage des relations fait de la question du contenu des voisins un problème de taille. Il suffit d'observer quelques-uns des couples rapprochés par l'ADA pour se rendre compte de l'hétérogénéité des relations extraites. Nous avons rapporté, dans le tableau 4.2, un échantillon de couples extraits des VDW porteurs de quelques-unes des relations évoquées dans les sections précédentes. La dernière colonne donne des exemples de contextes qui ont permis ces rapprochements.

Les couples qui y sont présentés illustrent certaines des spécificités de l'analyse fournie par la chaîne Syntex-Upery. On peut ainsi voir que les relations sont tout aussi bien portées par des paires d'arguments (*complet/total*) que de prédicats (*perdre*_SUJ*/gagner*_SUJ). Nous avons fait le choix de distinguer dans ce tableau trois grandes catégories de relations :

- les relations classiques ;
- les relations non classiques (à titre d'illustration, nous en avons ici sélectionné quatre) ;
- des couples qui ne portent pas de relation de sens manifeste.

Étant donnée la méthode mise en œuvre, on peut s'attendre à ce que les relations extraites soient principalement des relations classiques, qui sont de nature paradigmatique. On retrouve en effet parmi les voisins tout l'éventail des relations classiques que sont la synonymie, l'antonymie, l'hyperonymie, la co-hyponymie et la méronymie (la question de la proportion dans laquelle ces relations se manifestent dans les voisins sera abordée dans les sections suivantes).

Une certaine proportion des paires générées par cette méthode présentent des liens de sens manifestes sans pour autant qu'on puisse les caractériser en s'appuyant sur les typologies traditionnelles (*poète/poême*, *construction/outil*). Ces couples portent des relations non classiques (cf. section 4.1.1). Nous avons rangé parmi ces relations les liens qui relient des paires de mots appartenant à des catégories grammaticales différentes. Ces relations sont dites syntagmatiques. Nous avons en effet vu à la section 3.2.1 qu'Upéry permettait le rapprochement de paires intercatégorielles. Ces dernières représentent en moyenne 24 % des couples contenus dans chacune de nos trois

	relation	exemples de voisins	exemples de contextes
relations classiques	synonymie	*complet/total* *catastrophe/désastre*	*cécité*_MOD, *refonte*_MOD *ampleur*_DE, *tourner*_À, *éviter*_OBJ
	antonymie	*diurne/nocturne* *perdre*_SUJ/*gagner*_SUJ	*rapace*_MOD, *parade*_MOD *conservateur, Ajax, Arsenal*
	hyperonymie	*religion/islam* *artiste/peintre*	*précepte*_DE, *se convertir*_À *peindre*_SUJ, *inspiration*_POUR
	co-hyponymie	*Bruxelles/Vienne* *évêque/pape*	*palais*_À, *congrès*_À, *signer*_À *excommunier*_SUJ, *chapelain*_DE
	méronymie	*bras/doigt* *ville/quartier*	*amputation*_DE, *replier*_OBJ *construction*_DANS, *inaugurer*_DANS
relations non classiques	action/lieu	*étudier*_SUJ/*laboratoire*_DE *vendre*_OBJ/*magasin*_DE	*génétique, anthropologie, chimie* *accessoire, jouet, vêtement*
	action/agent	*colonisation*_MOD/*colon*_MOD *lire*_OBJ/*lecteur*_DE	*portugais, scandinave, romain* *cassette, cd, magazine, disque*
	producteur/produit	*poète/poème* *soleil/chaleur*	*anthologie*_DE, *citation*_DE *rayonnement*_DE, *exposition*_À
	action/instrument	*conquérir*_SUJ/*guerre*_CONTRE *construction/outil*	*Ottoman, Perse, royaume, pays* *géométrique*_MOD, *métallique*_MOD
autres associations		*mesurer*_OBJ/*favoriser*_SUJ *arbre/usine* *résider*_À/*incendier*_OBJ *antérieur*_À/*français*_DE *église/forêt* *théorie/paysage*	*inflation, humidité, taux, effet* *produire*_SUJ, *fournir*_SUJ *monastère, château, palais, hôtel* XVII^{e} *siècle,* XXIX^{e} *siècle* *renfermer*_SUJ, *incendie*_DE *exposer*_OBJ, *célèbre*_POUR

TAB. 4.2 – Exemples de relations sémantiques observables dans les VDW.

bases de voisins. La plupart d'entre elles – 80 % – sont des couples reliant un nom et un verbe. Certaines de ces paires sont morphologiquement liées (*lire*_OBJ/*lecteur*_DE), d'autres ne le sont pas (*vendre*_OBJ/*magasin*_DE, *conquérir*_SUJ/*guerre*_CONTRE). Sur cet aspect, les voisins se rapprochent du modèle *sac de mots*, qui tend à extraire ce type de relations thématiques.

Jusque-là, nous avons évoqué des cas où la proximité distributionnelle entre deux mots pouvait être identifiée soit comme l'une des relations en *-onymie*, soit comme une relation non classique de type *action/lieu* ou *producteur/produit* (certaines de ces relations sont recensées comme des fonctions lexicales dans la Théorie Sens-Texte). Ce n'est toutefois pas le cas de la plupart des couples rapprochés par l'ADA. Nous avons rapporté, à la ligne *Autres relations* du tableau 4.2, quelques couples dont l'interprétation est problématique. En effet, instinctivement, il est particulièrement difficile

d'expliquer leur rapprochement sans recourir au corpus. Par exemple, il n'y a aucune raison que le couple *usine/arbre* figure dans un thésaurus. Ces deux mots ne semblent pas partager suffisamment d'éléments de sens pour qu'un locuteur puisse les relier *via* une relation, qu'elle soit classique ou non classique. Une observation des contextes communs permet toutefois d'interpréter ce rapprochement comme de la co-hyponymie. Les deux mots apparaissent en effet en position sujet des verbes *produire* et *fournir*. Ils ont donc en commun le fait de mettre en œuvre des processus de génération (l'arbre produit des fruits, de l'oxygène, l'usine produit de l'électricité, des objets manufacturés, etc.). Il est intéressant de noter que ce chevauchement entre les propriétés de ces deux mots peut servir de support à des figures de style comme la métaphore : une requête sur Google nous permet de trouver de nombreux exemples de phrases comme "l'arbre est une usine d'épuration atmosphérique", "l'arbre fonctionne comme une usine biologique", "un arbre ressemble à une usine chimique géante", etc.

On peut considérer qu'on a ici affaire à un cas de relation construite en corpus. Le statut de ces relations dans notre ressource est problématique. D'un côté, elles sont tellement dépendantes du contexte qu'on peut difficilement les réexploiter, dans le cadre de la construction de thesaurus, par exemple. D'un autre côté, ces rapprochements sont révélateurs de fonctionnements en corpus qui sont tellement spécifiques qu'ils échappent à l'intuition : ils peuvent donc potentiellement nous permettre de mettre au jour des phénomènes inattendus.

Nous avons ici évoqué trois types de relations que la méthode distributionnelle – telle qu'elle a été mise en œuvre pour le développement des voisins – permet d'extraire :

- les relations classiques sont celles qui qui sont les mieux connues. Leur utilisation dans des applications de TAL représente un enjeu notoire. Ce sont des relations paradigmatiques et, en tant que telles, elles sont censées être extraites en priorité par un système comme Syntex-Upéry, qui s'appuie sur une modélisation syntaxique des contextes ;
- les relations non classiques présentent également un intérêt manifeste pour les systèmes de TAL – entre autres –, mais leur hétérogénéité peut constituer un frein à leur étude ainsi qu'à leur exploitation (Fabre, 2010) ;
- les relations qui ne peuvent être interprétées qu'en fonction du contexte (relations *ad hoc*) sont encore plus difficiles à appréhender, dans le sens où un locuteur *lambda* aurait du mal à les considérer comme des relations, en premier lieu. On aurait donc tendance à considérer ces paires comme étant du bruit. Toutefois, comme nous l'avons vu avec l'exemple de *arbre/usine*, un retour au corpus nous permet de trouver une explica-

tion à leur rapprochement, même si leur analyse n'est pas toujours aussi révélatrice. Par exemple, le couple *arbre/usine* se retrouve également dans les VDLM, mais il a été généré selon des modalités différentes : les deux mots ont été rapprochés à cause du fait qu'ils partagent de nombreux modifieurs génériques comme *gigantesque*, *vieux* ou *nombreux*, ce qui nous en apprend assez peu sur le fonctionnement de *arbre* et *usine* dans le corpus.

Ainsi, les relations contenues dans nos bases de voisins possèdent des statuts différents selon qu'elles sont établies comme des relations lexicales et largement étudiées comme telles, ou qu'elles résultent de fonctionnements plus locaux. Nous verrons que cette hétérogénéité va constituer une difficulté pour l'évaluation de ces ressources.

4.2 La problématique de l'évaluation

Nous avons évoqué à la section précédente les problèmes rencontrés lors de l'interprétation des résultats fournis par l'AD. En fait, la nature des rapprochements générés peut paraître si difficilement saisissable que toute tentative de les définir semblerait vouée à l'échec :

> It would seem that the information captured using [distributional methods] is not precisely syntactic, nor purely semantic – in some sense the only word that appears to fit is *distributional*. (Resnik, 1993, p. 18)

Toutefois, afin de faire une utilisation plus efficace des ressources lexicales acquises à partir de corpus, une étape d'évaluation s'avère nécessaire (Poibeau *et al.*, 2002; Poibeau et Messiant, 2008; Zargayouna et Nazarenko, 2010; Abbès *et al.*, 2011).

Il est difficile de définir de manière claire ce qui distingue une *bonne* paire de voisins d'une *mauvaise*. Nous avons évoqué à la section précédente la notion de *bruit*. En recherche d'information, le bruit désigne toute réponse à une requête jugée non pertinente. Cette notion dépend donc du jugement de l'utilisateur. De la même façon, la définition que nous donnons au bruit est entièrement liée au mode d'évaluation choisi : nous considérons comme du bruit tout couple de mots étant jugé non pertinent pour un mode d'évaluation donné.

En ce qui concerne les ressources acquises automatiquement à partir de corpus, la question de l'évaluation peut être abordée de différentes manières. La qualité des résultats produits pourra par exemple être mesurée à l'aune de leur potentiel à :

– améliorer les performances d'un système ;

– faciliter la construction d'une ontologie ;
– mimer les associations produites par des locuteurs.

Ces quelques exemples illustrent trois points de vue différents sur la question de l'évaluation. Ces derniers peuvent relever de deux démarches. On distingue en effet les modes d'évaluation selon qu'ils sont *intrinsèques* ou *extrinsèques*[2] (Sparck Jones et Galliers, 1996). Les méthodes extrinsèques, ou évaluations *par la tâche*, consistent à évaluer la ressource en fonction des performances du système dans lequel elle a été implémentée. Les méthodes intrinsèques évaluent la ressource en elle-même et pour elle-même. L'approche intrinsèque consiste à comparer les données obtenues automatiquement – dont la qualité est inconnue – à des ressources de référence (*gold standards*) construites manuellement comme WordNet ou EuroWordNet L'utilisation d'une ressource de référence est certainement le mode d'évaluation le plus répandu. Un autre type d'approche consiste à faire appel à l'intuition de locuteurs (spécialistes ou non-spécialistes).

Dans cette section, nous donnons une description de trois modes d'évaluation parmi les plus répandus en montrant dans quelle mesure ils sont applicables à nos données.

4.2.1 Utilisation de ressources de référence

Nous présentons ici deux grands types de données fréquemment utilisées comme des ressources de référence.

4.2.1.1 Réseaux lexicaux et dictionnaires

Nous avons vu, à la section 3.1, que pour ce qui est de l'évaluation de bases distributionnelles en anglais, les ressources qui ont été les plus utilisées sont WordNet et le Roget's thesaurus. Les équivalents de WordNet construits pour les langues européennes – EuroWordNet – ont également été utilisés comme des données de référence.

Ce mode d'évaluation pose toutefois certains problèmes :
– il implique tout d'abord l'existence d'une ressource de référence, ce qui n'est pas toujours le cas (Weeds, 2003) ;
– Curran (2004) identifie les trois plus grands défauts des ressources construites manuellement comme étant la variabilité de leur niveau de grain, leur couverture et leur cohérence ;
– le fait de construire des données manuellement est une démarche longue et fastidieuse qui implique de privilégier la qualité à la quantité. De

[2]Curran (2004) établit une distinction similaire entre les modes d'évaluation qu'il qualifie de *directs* et d'*indirects*.

fait, ces ressources peuvent pécher par la faiblesse de leur couverture (Peirsman *et al.*, 2008) ;
- ces ressources se limitent la plupart du temps à recenser des couples qui relèvent de relations lexicales classiques. Or, nous avons vu que les ressources distributionnelles faisaient émerger des relations qui vont au delà de la synonymie, de l'hyperonymie, etc.

Lorsqu'elles sont disponibles, ces ressources représentent malgré tout un moyen intéressant d'évaluer une base lexicale construite automatiquement puisque leur mode de constitution assure une qualité optimale des données *comparantes* (elles sont validées manuellement). Parmi les ressources lexicales disponibles pour le français, on peut citer WOLF (Fišer et Sagot, 2008), qui s'appuie sur le principe de WordNet, le dictionnaire de synonymes du CRISCO (Manguin, 2002; Manguin *et al.*, 2004), issu de la compilation des synonymes présents dans plusieurs dictionnaires de langue, ou encore le réseau JeuxDeMots (Lafourcade, 2007), construit de façon collaborative, qui possède l'avantage de recenser des couples appartenant à une grande variété de relations. Ces deux dernières bases lexicales sont utilisées dans la suite de notre étude. Nous les décrivons respectivement aux sections 4.3.2 et 4.3.3.

Le point commun entre toutes ces ressources est qu'elles n'ont pas été conçues dans le but de permettre l'évaluation de bases distributionnelles. Afin de disposer de données qui possèdent les propriétés optimales pour aborder cette tâche, Baroni et Lenci (2011) ont récemment développé BLESS, une base lexicale – en anglais – spécifiquement dédiée à l'évaluation des données acquises à l'aide de modèles distributionnels. Cette ressource est composée de 200 mots cibles (ou *concepts*) qui sont des noms concrets répartis équitablement dans 17 catégories, comme celle des fruits, des mammifères terrestres ou des outils. Les relations qui ont été choisies sont la co-hyponymie, l'hyperonymie, la méronymie, la relation d'attribution (le mot cible est relié à ses modifieurs les plus typiques : *fridge/empty*, *cathedral/baroque*, etc.) et la relation EVENT, qui associe un mot cible à un verbe qui se rapporte à une action, une activité ou un évènement qui lui est relatif (*toaster/burn*, *penguin/swim*, etc.). De plus, une relation RANDOM relie les mots cibles à des mots choisis de façon aléatoire. Les 26 554 couples que compte la base ont été construits en s'appuyant sur des ressources sémantiques (dont WordNet) et des corpus (Wikipédia et ukWaC). BLESS permettrait ainsi de caractériser la nature des associations générées par une base distributionnelle en permettant de mesurer la proportion de chacune des relations qui est captée. Les auteurs comparent ainsi plusieurs modèles *sac de mots* et montrent par exemple que, globalement, la co-hyponymie est la relation qui est la mieux captée, ou encore qu'une fenêtre de 2 mots ou de 20 ne change qu'assez peu la composition de la ressource générée (seule la proportion de couples

appartenant à la relation EVENT varie).

4.2.1.2 Données issues de la psycholinguistique

Nous classons dans cette catégorie les jeux de données collectés lors de tâches d'association (*priming*) ou d'évaluation de distance sémantique. Dans ces approches, le but est de trouver une corrélation entre les rapprochements extraits du corpus et les jugements émis par les locuteurs (Lindsey *et al.*, 2007; Wandmacher *et al.*, 2008) qui sont, de fait, considérés comme la référence. La plus connue de ces ressources est celle de Rubenstein et Goodenough (1965). Nous avons vu à la section 1.4 qu'elle consistait en un jeu de 65 couples de noms auxquels 51 juges humains ont attribué un score de similarité allant de 0 à 4 : *cord/smile* (0,02), *shore/voyage* (1,22), *brother/lad* (2,41), *magician/wizard* (3,21), *gem/jewel* (3,94). On peut également citer la ressource de Miller et Charles (1991), qui consiste en un sous-ensemble de 30 couples extraits du jeu de Rubenstein et Goodenough (1965), ou encore le jeu WordSimilarity–353 de Finkelstein *et al.* (2001), qui comporte 353 couples de mots. Ces données restent malgré tout assez rares et de taille limitée. Pour le français, on peut évoquer les travaux de Joubarne et Inkpen (2011) qui ont récemment traduit le jeu de Rubenstein et Goodenough (1965) et l'ont fait évaluer par des locuteurs francophones.

Les données issues de tests d'association ont été obtenues en demandant à des sujets de fournir le premier mot qui leur venait à l'esprit après que leur ait été présenté un mot stimulus. Les réponses sont classées en fonction du nombre de fois où elles sont produites. Par exemple, dans le jeu de Ferrand (2001), les réponses ont été les suivantes pour le stimulus *enfer* :

- paradis (30) ;
- diable (17) ;
- feu (9) ;
- rouge (6) ;
- damnation (6) ;
- chaud (5).

De telles données ont été construites – et sont librement téléchargeables, pour la plupart d'entre elles – pour une variété de mots comme les noms concrets (Alario et Ferrand, 1998), les mots abstraits (Ferrand, 2001), les verbes d'action (Duscherer et Mounoud, 2006), etc. À notre connaissance, ce type de données n'est pas utilisé pour l'évaluation de ressources acquises à partir de corpus.

4.2.2 Jugement humain et évaluation

Afin de pallier le manque de ressources de référence ou les limites de couverture qu'elles peuvent présenter, il est possible de faire appel au jugement linguistique de locuteurs. Cette méthode consiste à demander à des locuteurs d'exercer directement leur intuition sur un ensemble de paires extraites de la ressource à évaluer (soit les n meilleures paires, soit sur un échantillon prélevé de façon aléatoire (Evert et Krenn, 2001, 2005)). Cette approche introspective a l'avantage d'éviter les problèmes de couverture puisque tout locuteur est capable de juger de la proximité sémantique de deux mots, à partir du moment où leurs sens lui sont connus.

De plus, même si un tel protocole d'évaluation peut être relativement lourd à mettre en place et nécessiter des ressources pécuniaires qui ne sont pas forcément disponibles, il est toujours théoriquement possible de solliciter des locuteurs. Cette démarche a été facilitée ces dernières années par l'apparition de plate-formes de *crowdsourcing*[3] comme le Turc mécanique d'Amazon[4]. Ces dernières ont toutefois été la cible de critiques tant sur la qualité des données qu'elles permettent d'obtenir que sur l'éthique de leur fonctionnement (Adda *et al.*, 2011).

Cette tâche reste également sujette à discussion du fait qu'elle est assez peu naturelle et qu'elle implique la prise en compte de la variation inhérente au jugement humain (aussi bien *inter* qu'*intra*-annotateur) :

> An introspective approach is problematic, as the native speaker's personal knowledge or intuition is not directly accessible or observable. It cannot account for all possible contexts of a lexeme, nor trace all possible sense relation, and neither can it account for central and typical patterns of a paradigmatic term. [...] Introspection is, however, crucial for the interpretation of textual evidence, for the analysis of collocation results, and for the identification of lexical relations. (Storjohann, 2005, p. 8)

Le problème du contexte qui est évoqué ci-dessus a été étudié par Adam (2012) dans le cadre d'une tâche d'identification de relations parmi des couples extraits automatiquement. Ses résultats montrent que, lorsque les couples sont présentés aux annotateurs dans une phrase dans laquelle ils cooccurrent

[3] Jeff Howe et Mark Robinson du magazine *Wired*, qui sont à l'origine de ce terme, lui donnent la définition suivante : "Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call."

[4] `http://www.mturk.com`

(cf. exemple présenté à la section (4)) plutôt que hors contexte :
– l'accord inter-annotateurs passe de *faible/modéré* à *fort*[5] ;
– la proportion de couples identifiés comme étant sémantiquement pertinents passe de 36 % à 82,3 %.

4.2.3 Évaluation par la tâche

Comme son nom l'indique, le mode d'évaluation *par la tâche* consiste à évaluer les résultats d'une application qui s'appuie sur la ressource que l'on veut évaluer. On part du principe que les performances du système sont le reflet de la qualité de la ressource et, en cela, ce mode d'évaluation est qualifié d'*indirect* (Curran, 2004). Une telle démarche peut aller de soi lorsqu'une ressource est spécifiquement générée pour une application donnée, puisque la finalité est alors d'obtenir un système aux performances optimales et non de construire une ressource se conformant à un modèle théorique particulier, ou à celui du jugement humain.

La déclinaison la plus populaire de cette méthode repose sur l'utilisation de l'un des tests du TOEFL (*Test of English as a Foreign Language*) qui consiste à sélectionner le synonyme d'un mot donné parmi quatre propositions. Il est intéressant de noter que le modèle développé par Bullinaria et Levy (2012) a récemment obtenu le score de 100 % pour cette tâche[6].

Ce type d'évaluation a également été utilisé dans des applications bien connues du TAL comme la recherche d'information (Ruge, 1992), la désambiguïsation (Weeds et Weir, 2005), les systèmes de question-réponse (van der Plas, 2008) ou la segmentation thématique (Adam, 2012). La nature même de ce mode d'évaluation implique qu'il peut se décliner de diverses manières. On peut ainsi citer les travaux de Budanitsky et Hirst (2006), qui ont comparé les performances de cinq ressources générées avec des mesures différentes sur une tâche de détection de *malapropisms*[7].

Cette méthode possède l'avantage de pouvoir constituer une alternative à l'indisponibilité de ressource de référence. Elle pèche toutefois par son aspect *boîte noire* : dans la mesure où le seul paramètre observable est la variation de performance du système, on ne peut faire que des suppositions sur les aspects de la ressource qui ont influé sur ses performances. En cela, l'approche par la tâche empêche toute démarche interprétative et ne constitue donc pas une

[5]Selon l'échelle définie par Landis et Koch (1977).

[6]Se référer au wiki de l'Association for Computational Linguistics (`http://aclweb.org/aclwiki`) pour un état de l'art des performances qui ont pu être enregistrées pour cette tâche.

[7]Ce terme désigne les cas où un mot est remplacé par erreur par l'un de ses paronymes (*conjoncture* à la place de *conjecture*, par exemple).

méthode d'évaluation satisfaisante à nos yeux.

4.3 Mesurer le recouvrement entre les voisins et des lexiques externes

Nous avons évoqué à la section précédente les approches les plus répandues pour l'évaluation de ressources lexicales acquises automatiquement à partir de corpus. Parmi ces approches, la comparaison à des dictionnaires ou réseaux lexicaux est celle qui nous paraît la plus adaptée à notre conception de l'évaluation des voisins. En effet, cette approche qui consiste à mettre en parallèle un lexique de synonymes, antonymes, etc. et les voisins nous permet, dans un premier temps, d'avoir un aperçu global du contenu des bases distributionnelles puis, dans un deuxième temps, d'adopter un point de vue plus local pour isoler certains phénomènes linguistiques qui sont à l'œuvre dans les corpus. Dans cette section, nous abordons la première de ces deux approches : après avoir opéré un prétraitement des données (4.3.1), nous comparons successivement les voisins de Wikipédia (VDW), de Le Monde (VDLM) et de Frantext (VDF) à un dictionnaire de synonymes – le DES – (section 4.3.2) et à un réseau lexical contenant une grande variété de relations – JeuxDeMots (JDM) – (section 4.3.3).

La raison pour laquelle nous avons choisi de travailler avec ces deux types de ressources repose en premier lieu sur un critère de disponibilité. En effet, il nous a été impossible de disposer d'une ressource comme la version d'EuroWordNet. Le choix de réseaux lexicaux/thesaurus/dictionnaires étant limité pour le français, nous avons choisi de réutiliser la version du DES disponible au CLLE-ERSS – dont Bourigault et Galy (2005) se sont déjà servi dans le cadre de l'évaluation des voisins – et la ressource JeuxDeMots, qui possède l'avantage d'être librement téléchargeable. Nous discutons plus loin des conséquences de ces choix.

4.3.1 Prétraitements

Avant de comparer les ressources, nous avons procédé à une étape d'homogénéisation. La nécessité d'adapter les données à évaluer au format de la ressource de référence constitue l'un des inconvénients de ce mode d'évaluation (Poibeau et Messiant, 2008). En effet, alors que les lexiques contiennent des couples de mots, les voisins rapprochent des prédicats et des arguments. Pour les besoins de la comparaison, nous avons donc dû – temporairement – mettre de côté la relation contenue dans chaque prédicat. On perd alors une information de première importance sur la nature des couples de prédicats et

	avant dédoublonnage	après dédoublonnage
VDW	3 922 657	2 739 761
VDLM	5 525 480	4 103 968
VDF	792 356	673 804

TAB. 4.3 – Résultat du dédoublonnage des voisins.

d'arguments qui ont été rapprochés : alors que le dictionnaire de synonymes que nous utilisons rapproche *souligner* et *insister*, les voisins rapprochent *souligner*_OBJ et *insister*_SUR (Bourigault et Galy, 2005). La nature de la relation portée par les prédicats sera toutefois prise en considération lors de la phase d'interprétation des résultats.

Cette différence entraîne un phénomène de redondance dans les couples de voisins, puisque de nombreux couples de prédicats relient deux lemmes identiques *via* des relations différentes. C'est notamment le cas du couple de verbes *se situer/s'étendre* qui, dans les VDW, apparaît dans 28 couples de prédicats différents :

- *se situer*_DANS/*s'étendre*_DE
- *se situer*_ENTRE/*s'étendre*_DANS
- *se situer*_SUR/*s'étendre*_LE LONG DE
- etc.

Étant donnée l'importance de ce phénomène – certains couples présentent jusqu'à 70 variantes –, nous avons choisi de dédoublonner les trois bases de voisins afin d'obtenir des mesures de recouvrement les plus fidèles possibles (le score de Lin le plus élevé de toutes les variantes étant celui qui a été conservé pour les couples uniques). Comme on peut le voir au tableau 4.3, ce dédoublonnage fait considérablement baisser le nombre de couples de chaque base de voisins (de 24 % en moyenne).

4.3.2 Le Dictionnaire électronique des synonymes

Dans cette section, nous présentons dans un premier temps les propriétés de la ressource avant de décrire la façon dont nous l'avons comparée aux voisins, ainsi que les résultats que nous avons obtenus. Le calcul du recouvrement entre les trois bases de voisins et le DES a été fait selon deux modalités. La première a consisté à mesurer consécutivement le nombre de couples communs entre les trois bases de voisins – dans leur intégralité – et le DES. Nous verrons que les résultats obtenus nous ont incité à adopter une deuxième approche du recouvrement dans laquelle seul le lexique partagé a

	% des voisins		% du DES	
	Lexique intégral	Lexique partagé	Lexique intégral	Lexique partagé
VDW	1,5	1,8	11	41,6
VDLM	1,1	1,5	12,4	33,7
VDF	2,8	3,6	4,9	21,1

Tab. 4.4 – Proportion du recouvrement entre les trois bases de voisins et le DES.

été pris en compte.

4.3.2.1 Présentation de la ressource

Le Dictionnaire électronique des synonymes (Manguin, 2002; Manguin *et al.*, 2004) – DES, ou encore *Dicosyn* – a été développé au sein du laboratoire CRISCO, de l'Université de Caen. Il regroupe les synonymes relevés dans sept dictionnaires (dictionnaires analogiques et dictionnaires de synonymes) – le *Bailly*, le *Benac*, le *Du Chazaud*, le *Guizot*, le *Lafaye*, le *Larousse* et le *Robert* – et est consultable à l'adresse suivante : `http://www.crisco.unicaen.fr/des/`. La version dont nous disposons comporte 389 182 paires de noms (52 %), verbes (26 %) et adjectifs (22 %)[8], mots simples ou syntagmes Il est à noter que les paires ont été symétrisées (pour une relation *A/B* a été générée une relation *B/A*) : le dictionnaire compte donc 194 576 relations de synonymie réciproques. Le DES étant régulièrement mis à jour, il comptait, au 25 octobre 2012, 200 849 relations réciproques. Les taux de recouvrement obtenus, qui sont commentés par la suite, sont rapportés au tableau 4.4 et illustrés à la figure 4.1.

4.3.2.2 Comparaison des ressources intégrales

Précision La proportion de synonymes du DES dans les voisins varie de 1,1 % à 2,8 % (ces proportions sont de l'ordre de celles qui ont été obtenues par Bourigault et Galy (2005) avec les VDLM). Cela correspond à 41 718 couples dans les VDW, 46 942 dans les VDLM et 18 590 dans les VDF. Le fait que cette proportion soit la plus élevée dans les VDF peut s'expliquer par la différence de taille entre les VDF et les deux autres bases de voisins.

[8]Ces proportions s'appuient sur un étiquetage automatique accompagné d'une vérification manuelle effectués par Mai Ho-Dac et Franck Sajous (CLLE-ERSS).

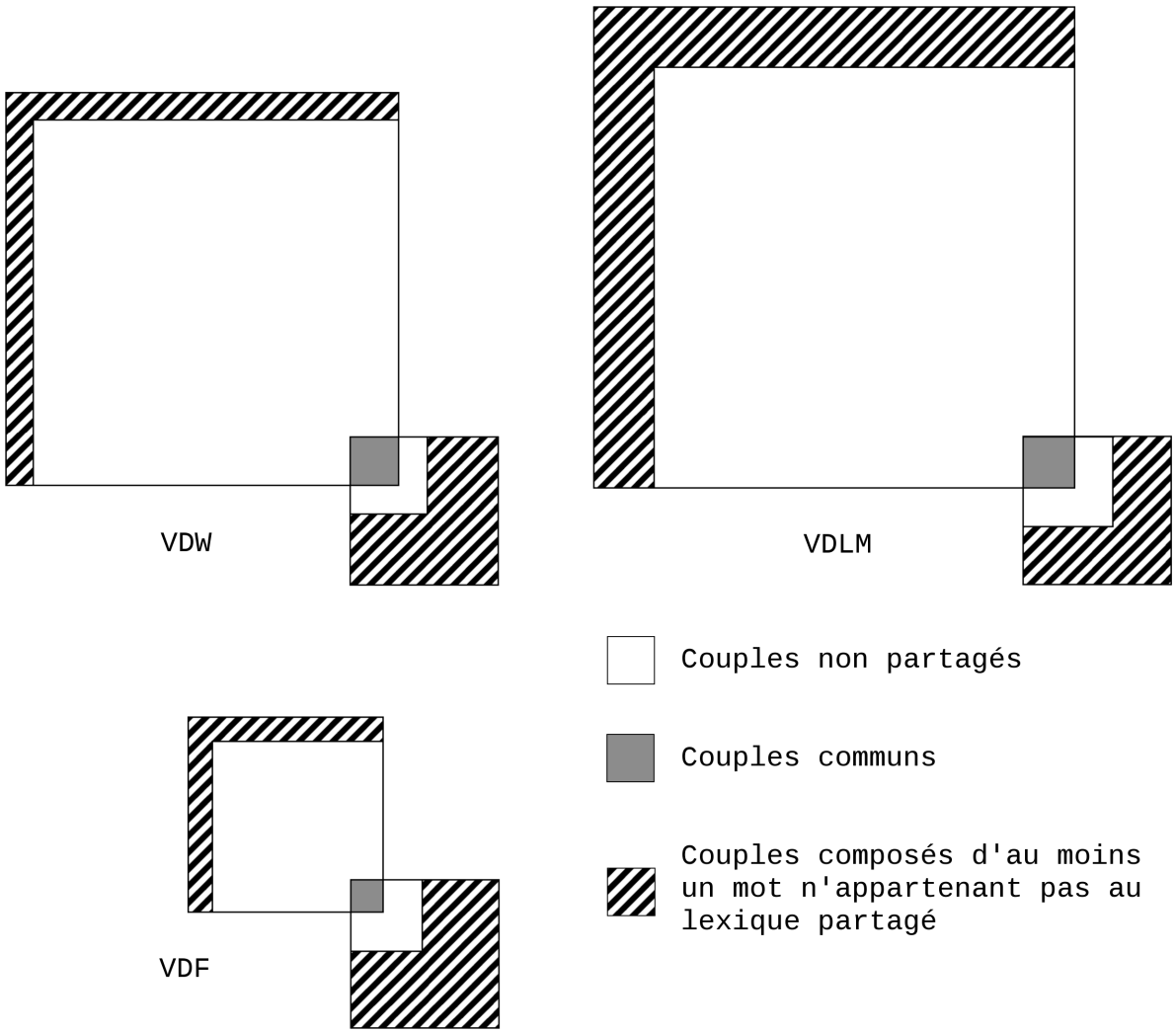


FIG. 4.1 – Illustration du recouvrement entre les trois bases de voisins et le DES. Le carré de gauche représente les VDW/VDLM/VDF, celui de droite représente le DES.

On peut également soupçonner une influence favorable de la nature littéraire du corpus à partir duquel ont été générés les VDF.

Rappel Réciproquement, entre 4,9 % et 12,4 % des synonymes du DES ont été captés par l'ADA (l'équivalent en nombre de couples a été donné dans le paragraphe précédent). Ici aussi, les proportions observées sont à peu près égales pour les VDW et les VDLM, qui permettent respectivement de capter 11 % et 12,4 % des synonymes recensés dans le DES. La différence de taille entre les bases de voisins se fait encore une fois ressentir : alors que la taille réduite des VDF leur permettait d'obtenir un degré de recouvrement avec le DES légèrement supérieur à celui des VDW/VDLM, on voit que seulement 4,9 % des synonymes du DES ont été captés par les VDF. Il s'agit ici d'un rapport entre les notions de *précision* et de *rappel*, sur lesquelles nous aurons l'occasion de revenir par la suite.

La faiblesse générale de ces taux a de quoi surprendre. Il paraît en effet étonnant de voir que le principe distributionnel ne permet de capter, en moyenne, que moins de 10 % des synonymes du DES. Nous allons voir, toutefois, que ces proportions sont à nuancer.

4.3.2.3 Comparaison des ressources à lexique partagé

Une des raisons pouvant expliquer l'absence d'une paire de synonymes du DES dans les voisins – et *vice versa* – est la différence de couverture lexicale. Il convient en effet de distinguer les raisons liées à l'absence d'un certain lexique dans les deux ressources de celles qui relèvent de la distribution des mots du couple dans le corpus qui a servi à générer la base. Par exemple, le couple *bottine/soulier* est absent des VDW car *bottine* et *soulier* sont très rares dans le corpus Wikipédia : ils n'apparaissent dans aucune paire de voisins. Il s'agit ici d'un problème lié à la couverture lexicale. Ce cas de figure est à distinguer de celui du couple de synonymes *aborder/attaquer* : les deux termes de la paire sont bien présents, séparément, dans les VDW mais leurs distributions respectives ne sont pas assez similaires pour qu'ils y apparaissent en tant que couple (contrairement, à *aborder/traiter* et *aborder/évoquer*, qui relèvent d'une autre acception du verbe plus répandue dans le corpus).

Afin d'avoir une estimation du recouvrement des ressources moins biaisée par la différence de couverture lexicale, nous avons choisi de recalculer le recouvrement en écartant toute paire de voisins dont au moins un des membres n'apparaît pas dans le DES (et *vice versa*). Ces paires apparaissent hachurées à la figure 4.1.

Précision Le fait d'exclure le lexique non partagé a une influence variable sur la taille des bases distributionnelles. En effet, la proportion de couples composés d'au moins un mot qui n'apparaît pas dans le DES est de 13 % dans les VDW, de 24 % dans les VDLM et de 23 % dans les VDF. Toutefois, comme on peut le voir au tableau 4.4, cela n'a que peu d'influence sur la proportion de synonymes. Le nombre de couples reste en effet considérable et le pourcentage de synonymes dans chacune des ressources n'augmente que de 0,3 % dans les VDW, de 0,4 % dans les VDLM et de 0,8 % dans les VDF.

Rappel En revanche, on peut voir que les proportions de synonymes captées sont, en moyenne, multipliées par 3,6. Cela est dû au fait que la non-prise en compte des couples du lexique non partagé réduit la taille du DES de 71 % (en moyenne). La proportion de synonymes captés par les VDW atteint ainsi les 41,6 %.

Loin de constituer une étape triviale du protocole de mesure du recouvrement, cette manipulation nous a permis de mettre en lumière un décalage flagrant entre nos voisins – les ressources *comparées* – et le DES – la ressource *comparante* : le fait de savoir que les trois quarts des couples du DES sont impossibles à trouver dans les VDW parce qu'ils contiennent au moins un mot qui est très rare dans le corpus Wikipédia – voire absent – permet de juger la base de voisins à sa juste valeur.

Analyse des résultats On peut évoquer plusieurs raisons expliquant cette différence entre les couvertures lexicales du DES et des voisins. Ainsi, comme l'on considère habituellement que les noms propres ne peuvent entretenir de relations de synonymie, le DES n'en contient aucun (si l'on met de côté les locutions comme *vieille France*). En moyenne, un peu plus de 10 % de voisins en contiennent au moins un.

De plus, puisque la relation de synonymie ne se manifeste qu'entre deux mots appartenant à la même catégorie grammaticale, le DES ne contient aucune paire intercatégorielle, alors que nous avons vu qu'en moyenne, presque un quart des voisins sont des couples intercatégoriels.

On peut également évoquer le problème des unités polylexicales. Syntex lemmatise systématiquement chacun des éléments des syntagmes nominaux : *affaires étrangères* est présent dans les voisins sous la forme *affaire étranger*, de la même façon que *extrême droite* et *cahier des charges* le sont sous les formes *extrême droit* et *cahier de charge*. Étant donnée la très faible proportion de syntagmes nominaux parmi les voisins, nous n'avons pas jugé utile de mettre en place une procédure visant à rétablir leur forme *usuelle*. Ainsi, les 8 % de couples du DES composés d'au moins une unité complexe auront

très peu de chance d'être présents dans les bases de voisins.

4.3.3 JeuxDeMots

Comme nous l'avons fait pour le DES, nous faisons une description du mode de constitution et du contenu de JeuxDeMots avant de comparer cette ressource avec les voisins.

4.3.3.1 Présentation de la ressource

La deuxième ressource à laquelle nous comparons les voisins est le réseau lexical JeuxDeMots (Lafourcade, 2007) – JDM –, développé au LIRMM (Université de Montpellier). Il s'agit d'une ressource issue du *crowdsourcing* : elle est construite de façon collaborative par des agents humains – experts et non experts confondus – *via* une application en ligne. Nous avons évoqué à la section 4.2.2 quelques-unes des critiques qui ont pu être adressées à l'encontre du *crowdsourcing*. La différence entre JeuxDeMots et le Turc mécanique réside dans le fait que les contributeurs de JeuxDeMots ne sont pas rémunérés : l'enrichissement du réseau repose sur un jeu en ligne (accessible à l'adresse suivante : `http://www.lirmm.fr/jeuxdemots/`) consistant à proposer une série de mots pour un mot cible et une relation donnés[9]. Les consignes sont formulées de façon à être compréhensibles pour les joueurs n'ayant pas de connaissances en linguistique :

- "Donner des GÉNÉRIQUES pour le terme qui suit (par exemple, *véhicule* pour *voiture*, *félin*, *animal* pour *chat*)" pour le relation d'hyperonymie ;
- "Donner des PARTIES du terme suivant : (une partie est une composante de l'objet, par exemple : *moteur*, *roue*, etc. pour *voiture* – ou encore *couverture*, *pages*, *chapitre* etc. pour *livre*)" pour la relation de méronymie.

Le joueur a une minute pour donner le plus de réponses possibles (cf. figure 4.2). Passé ce délai, ses réponses sont comparées aux réponses qu'ont fournies les joueurs précédents pour la même partie (cf. figure 4.3). Si c'est la première fois que la partie est proposée, alors les résultats du joueur sont enregistrés en attendant que la partie soit proposée à un autre joueur. Les réponses communes à deux joueurs sont ajoutées au réseau. Si le lien était déjà présent, alors il est renforcé selon un système de pondération. Ainsi, dans le cas de figure illustré à la figure 4.3, les liens *table*–MÉRONYME–*salon*, *fauteuil*–MÉRONYME–*salon*, *tapis*–MÉRONYME–*salon* et *canapé*–MÉRONYME–*salon* ont été consolidés dans le réseau. Pour chaque réponse commune, les joueurs sont

[9]Voir Lafourcade (2007) pour une description plus détaillée du protocole.

FIG. 4.2 – Interface de JeuxDeMots : phase de jeu.

FIG. 4.3 – Interface de JeuxDeMots : comparaison des résultats (nous avons encadré en rouge les réponses communes aux deux joueurs).

récompensés par des points et des crédits qui leur permettent de débloquer certaines fonctionnalités du jeu. Les données générées peuvent être téléchargées librement[10].

La fiabilité des données est garantie par le fait que le réseau ne contient que des couples de mots qui ont été fournis par deux joueurs différents (un dispositif permet d'empêcher qu'un joueur ne retombe deux fois sur une même partie et ne valide ainsi ses propres réponses). Toutefois, nous verrons que les couples que contient JeuxDeMots sont fortement influencés par les processus de jeu : le but des joueurs est en effet d'obtenir le maximum de points et non de fournir les réponses les plus pertinentes du point de vue lexicographique (cf. sections 7.2.1 et 8.3.1). De ce fait, de nombreux couples contenus dans cette ressource paraissent assez peu intuitifs. Il est difficile d'avoir un ordre d'idée de la proportion que représentent ce type de couples (ce qui consisterait à évaluer la ressource évaluante...). Toutefois, étant donné que, comme nous l'avons souligné plus tôt, les ressources contenant des relations qui vont au delà de la synonymie sont quasiment inexistantes (ou indisponibles) pour le français, nous avons fait le choix de nous servir malgré tout de JeuxDeMots comme ressource comparante en assumant ses imperfections.

L'un des intérêts de cette ressource est en effet qu'elle contient des relations aussi bien classiques que non classiques. Nous avons rapporté au tableau 4.5 l'éventail des relations qui sont proposées aux joueurs ainsi que le nombre de couples qu'elles relient dans le réseau[11] dans sa dernière version (c'est-à-dire, à l'heure où nous écrivons ces lignes, celle du 5 février 2013). On peut constater que, malgré une certaine variété dans la nature des relations proposées, environ 82 % des 1 429 193 couples sont reliés par l'une des 4 relations les plus fréquentes. La relation IDÉE ASSOCIÉE relie à elle seule plus de la moitié des couples, ce qui est dommageable dans le sens où cette relation est particulièrement vague et qu'elle peut englober la plupart des autres relations (deux synonymes renvoient à des idées associées par une relation d'identité, etc.).

4.3.3.2 Mesure du recouvrement

Étant donné que JDM contient 24 relations de natures différentes, comparer l'intégralité de la ressource avec les voisins, comme nous l'avons fait avec le DES, n'aurait pas eu beaucoup de sens. Nous avons donc choisi de considérer JDM comme un ensemble de 24 ressources lexicales contenant chacune

[10] `http://www.lirmm.fr/~lafourcade/JDM-LEXICALNET-FR/`

[11] Dans sa version téléchargeable, le réseau contient d'autres relations qui ne sont pas proposées aux joueurs ainsi que des relations internes qui relient un mot à son lemme, sa classe grammaticale, etc.

fréquence	%	nom	description	exemples
820 039	57,4	r_associated	idée associée à m	*aventure/voyage, thérapie/soigner*
165 005	11,5	r_syn	synonyme de m	*casse/destruction, déviation/écart*
93 168	6,5	r_isa	hyperonyme de m	*bricolage/activité, coupe/récompense*
92 598	6,5	r_domain	domaine relatif à m	*guitare/musique, moelle/boucherie*
27 695	1,9	r_lieu	lieu où peut se trouver m	*magicien/scène, chenille/forêt*
26 013	1,8	r_carac	propriété de m	*mannequin/élancée, verre/cassé*
23 641	1,7	r_has_part	méronyme de m	*deltaplane/toile, remorque/roue*
19 245	1,3	r_hypo	hyponyme de m	*salade/scarole, ministre/Fillon*
19 131	1,3	r_agent	entité qui effectue l'action m	*marteler/artisan, verser/serveur*
17 943	1,3	r_locution	locution contenant m	*pays/pays natal, mort/tête de mort*
15 126	1,1	r_meaning	sens de m	*air/attitude, air/musique*
14 421	1	r_patient	entité qui subit l'action m	*épouser/femme, clore/discussion*
14 255	1	r_anto	antonyme de m	*manque/opulence, amorti/amplifié*
14 081	1	r_familly	famille morphologique de m	*terrorisme/terreur, élastique/élasticité*
13 488	0,9	r_holo	holonyme de m	*laine/écharpe, nerf/corps*
10 535	0,7	r_instr	objet qui réalise l'action m	*dessiner/crayon, remplir/cuillère*
8142	0,6	r_causatif	cause de m	*casser/colère, désert/aridité*
6833	0,5	r_conseq	conséquence de m	*élection/président, clignotant/tourner*
6037	0,4	r_maner	manière d'effectuer l'action m	*refuser/poliment, cuire/trop*
5531	0,4	r_sentiment	sentiment que suscite m	*pollution/peur, internet/bonheur*
4956	0,3	r_magn	intensification de m	*fleur/bouquet, appartement/château*
4833	0,3	r_antimagn	affaiblissement de m	*épée/couteau, mort/maladie*
4097	0,3	r_telic_role	but/fonction de m	*message/informer, antonymie/opposer*
2380	0,2	r_agentif_role	mode de création de m	*orange/cueillir, livre/imprimer*

Tab. 4.5 – Répartition des relations proposées dans JDM. Dans les couples donnés en exemple, le mot cible m apparaît toujours en première position.

	VDW			VDLM			VDF		
	brut	% vois.	% rel.	brut	% vois.	% rel.	brut	% vois.	% rel.
r_associated	46 678	1,9	31,8	50 670	1,5	26,8	17 583	3,2	15,6
r_syn	19 392	0,9	46,5	22 286	0,8	39,2	8973	1,8	26,4
r_isa	2660	0,3	38,9	2264	0,3	32,4	1269	0,7	25,3
r_domain	2589	0,2	14,8	2147	0,1	12,7	353	0,1	5,3
r_lieu	2290	0,7	33,7	2084	0,7	29,4	822	1,0	13,9
r_carac	577	0,2	11,4	550	0,2	10,4	335	0,4	7,3
r_has_part	1769	0,5	29,6	1389	0,4	23,9	967	1,0	18,8
r_hypo	1636	0,4	45,4	1469	0,4	39,8	1013	0,9	37,7
r_agent	639	0,3	16,3	787	0,2	16,2	196	0,2	5,0
r_locution	93	0,0	30,0	85	0,0	18,8	72	0,1	59,0
r_meaning	2930	0,3	49,6	2842	0,2	44,2	1461	0,6	29,8
r_patient	742	0,2	18,5	859	0,2	17,3	241	0,2	5,7
r_anto	1194	0,3	38,9	1430	0,3	33,2	731	0,7	26,2
r_familly	877	0,2	28,7	890	0,1	25,5	397	0,3	19,5
r_holo	1725	0,5	36,9	1506	0,5	32,9	912	1,1	22,8
r_instr	300	0,2	16,2	360	0,2	15,6	193	0,3	9,9
r_causatif	616	0,2	33,2	677	0,2	30,6	262	0,4	17,4
r_conseq	559	0,2	36,0	557	0,2	31,3	201	0,3	15,7
r_maner	8	0,2	7,1	5	0,1	3,9	9	0,2	7,9
r_sentiment	347	0,4	17,0	455	0,4	17,8	214	0,6	9,9
r_magn	942	0,4	59,3	871	0,4	48,9	646	1,0	45,9
r_antimagn	991	0,4	57,5	932	0,4	49,8	771	0,9	46,6
r_telic_role	120	0,1	13,5	131	0,1	12,7	30	0,1	4,8
r_agentif_role	132	0,2	16,0	125	0,2	14,0	30	0,1	5,1

TAB. 4.6 – Recouvrement entre les voisins et les relations de JDM.

une relation donnée et de les comparer individuellement à nos trois bases de voisins. Les résultats sont présentés au tableau 4.6.

Précision Le nombre de couples communs varie énormément en fonction de la fréquence de la relation. Le fait que les couples portant la relation r_maner ne soient pratiquement pas repérés s'explique par le fait que cette relation relie – en principe – un verbe et un adverbe, et que les adverbes ne sont pas pris en compte lors du calcul distributionnel. La relation r_locution est également très peu trouvée dans les voisins. Cela peut s'expliquer par le fait que les locutions en question peuvent être des syntagmes verbaux (*train/être en train de*), adjectivaux (*chou/bête comme chou*), adverbiaux (*feu/tout feu tout flamme*), voire des phrases entières (*curiosité/la curiosité est un vilain défaut*). Or, les voisins ne sont constitués que de noms, verbes et adjectifs simples, éventuellement de syntagmes nominaux, mais en très faible quantité.

Les chiffres donnés dans les colonnes *% vois.* indiquent la proportion de

voisins couverts par chacune des relations. Il est à noter que les calculs ont été directement faits en ne prenant en compte que les couples appartenant au lexique commun. De plus, puisque les recouvrements ont été calculés de façon indépendante pour chacune des relations, le nombre de voisins pris en compte varie pour chaque relation :

- les 577 couples de la relation r_carac représentent 0,2 % des 342 452 VDW considérés ;
- les 132 couples de la relation r_agentif_role représentent 0,2 % des 65 465 VDW considérés.

Étant donné que la grande majorité des relations ne relient qu'un faible nombre de couples – comparativement au DES, par exemple –, le pourcentage de voisins qu'elles recouvrent est également très faible. Ainsi, seule la relation r_associated est détectée dans plus de 1 % des VDW et des VDLM. Ces scores sont plus élevés dans les VDF, qui sont d'une taille plus réduite.

Rappel Les colonnes *% rel.* indiquent la proportion de couples appartenant à la relation concernée qui sont captés par les voisins. Les relations suivantes sont celles qui sont les mieux repérées :

- les relations r_magn et r_antimagn sont toutes les deux repérées à hauteur de 51,3 % (toutes bases de voisins confondues). La consigne donnée pour ces relations est "Qu'est ce qui est PLUS/MOINS INTENSE que le terme qui suit". Parmi les couples produits par les joueurs – et repérés dans les VDW –, on trouve :
  - *major/lieutenant*, *baguette/bâton* ou *peur/terreur* pour la relation r_magn ;
  - *passion/amitié*, *ennemi/concurrent* ou *fleuve/rivière* pour la relation r_antimagn.

  On peut expliquer la raison pour laquelle ces couples ont été si bien repérés par le fait qu'ils relient deux mots qui sont sémantiquement similaires : *peur* et *terreur* sont des sentiments (négatifs), *fleuve* et *rivière* renvoient à des cours d'eau, etc.
- la relation r_meaning a été captée, en moyenne, à hauteur de 41,2%. La consigne qui correspond à cette relation est particulièrement floue : "Quels SENS/SIGNIFICATIONS pouvez vous donner au terme qui suit (il s'agira de termes évoquant chacun des sens possibles [...])". Il en résulte que les couples liés par cette relation sont de nature hétérogène. La plupart d'entre eux sont composés de synonymes (*interpréter/analyser*, *interpréter/jouer*), d'hyperonymes (*gaz/énergie*), de cohyponymes (*loup/chacal*) ou encore de relations types syntagmatiques (*tondre/gazon*).

– vient ensuite la relation r_hypo, qui a été repérée dans des proportions similaires (42 %).

Parmi les relations les moins bien captées, on trouve les relations suivantes :

– la relation r_maner (qui relie un verbe à un adverbe) ;
– les couples reliés par la relation r_carac sont des couples nom/adjectif ; Or, ces couples n'ont que très peu de chances d'être repérés dans l'analyse (ils ne représentent, en moyenne, que 0,4 % des voisins).
– vient ensuite la relation r_telic_role qui n'a été captée qu'à 10,3 %. Cette proportion est semblable à celles des autres relations nom/verbe, à savoir r_agentif_role (11,7 %), r_agent (12,5 %), r_patient (13,8 %) et r_instr (13,9 %) ;
– la relation r_domain n'a été repérée qu'à 10,9 %. Elle relie des couples comme *frégate/navigation*, *équipe/sport* ou *pompe/mécanique*. Ces couples expriment des relations associatives dont on sait qu'elles sont mieux repérées par les approches *sac de mots*.

Ces résultats – ainsi que ceux que nous avons obtenus avec le DES – nous permettent de mettre en exergue les limites de l'approche consistant à comparer des données acquises automatiquement à des ressources de référence. Dans la section suivante, nous poursuivons cette démarche à la lumière des observations que nous avons pu faire durant cette étape : nous mettons notamment en place un protocole de mesure de recouvrement qui se veut moins biaisé par les effets de fréquence en nous focalisant sur un nombre plus restreint de relations.

4.4 Critères influençant la composition des voisins

Dans cette section, nous nous livrons à plusieurs mesures de recouvrement en adoptant un protocole inspiré de celui de van der Plas (2008). Le but est d'observer l'influence de certaines propriétés des voisins sur leur propension à capter certaines relations lexicales. Ces critères sont les suivants :

– leur fréquence (section 4.4.1) ;
– leur catégorie grammaticale (section 4.4.2) ;
– leur statut d'argument ou de prédicat (section 4.4.3) ;
– le corpus à partir duquel ils ont été calculés (section 4.4.4).

Le protocole consiste à :

1. constituer plusieurs séries de 1000 mots cibles en fonction de ces critères ;
2. comparer nos ressources de référence aux n premiers voisins de chacun

de ces mots (classés par Lin décroissant) n valant successivement 1, 5 et 10.

Nous avons ici effectué ces mesures sur les VDW.

Nous avons choisi de nous focaliser sur les relations qui sont étudiées dans les chapitres 5 à 8, à savoir les relations classiques de sémantique lexicale que sont la synonymie, l'antonymie, l'hyperonymie, et la méronymie. Afin d'homogénéiser les ressources comparées, nous avons symétrisé les relations de JDM (les couples de voisins sont déjà symétriques). C'est-à-dire que pour chaque couple A/B d'une relation donnée, nous avons généré un couple B/A. Dans le cas de l'hyperonymie et de la méronymie cette symétrisation donnait lieu à des redondances. Nous avons donc mélangé, d'une part, les hyponymes avec les hyperonymes et, d'autre part, les méronymes avec les holonymes. Nous disposons donc pour ces expériences de quatre groupes de relations :

- les 389 182 synonymes du DES (ils ont été préférés à ceux de JDM, dont la quantité est deux fois moindre) ;
- 150 451 hyponymes et hyperonymes ;
- 48 815 méronymes et holonymes ;
- 22 922 antonymes.

4.4.1 La fréquence

Dans un premier temps, nous avons mené une étude consistant à extraire trois groupes de mots cibles dans des tranches de fréquence différentes (calculées à partir du corpus Wikipédia). Nous avons ainsi choisi de comparer :

- les 1000 mots les plus fréquents (de 8471 à 237 374 occurrences) ;
- 1000 mots dont les fréquences se situent aux alentours de la médiane (de 620 à 774 occurrences) ;
- les 1000 mots les moins fréquents (de 30 à 124 occurrences).

Il est à noter que le fait que chaque mot doive disposer d'au moins 10 voisins limite la possibilité de choisir des mots très rares : moins un mot est fréquent dans le corpus, moins il y a de chances que sa distribution soit rapprochée de celle des autres mots du corpus.

Les résultats rapportés au tableau 4.7 montrent que la proportion de relations lexicales parmi les voisins baisse à mesure que la fréquence des mots cibles décroît. Ce phénomène s'observe pour toutes les relations, quelle que soit la valeur de n.

On voit également que la relation de synonymie est dans tous les cas la plus représentée dans les VDW, suivie de la relation d'hypo/hyperonymie. Il est toutefois difficile de savoir dans quelle mesure les proportions calculées pour ces deux relations sont influencées par la différence entre le nombre de

	fréq. élevées			fréq. médianes			fréq. faibles		
	n=1	n=5	n=10	n=1	n=5	n=10	n=1	n=5	n=10
synonymie (DES)	19,7	13	10,6	14,9	9	6,8	7	4,4	3,9
antonymie	3,9	2	1,3	1	0,6	0,4	0,5	0,2	0,2
hyperonymie	9,7	6,1	4,8	3,8	1,9	1,4	0,2	0,2	0,1
méronymie	3,3	2,7	2,2	0,4	0,4	0,4	0	0,1	0,1
total	36,6	23,8	18,9	20,1	11,9	9	7,7	4,9	4,3

TAB. 4.7 – Recouvrement entre les voisins de différentes fréquences et nos ressources de référence (en %).

synonymes et le nombre d'hypo/hyperonymes contenus dans nos ressources de référence. On constate malgré tout que les proportions de méro/holonymes et d'antonymes sont assez semblables alors que JDM contient deux fois plus de méro/holonymes que d'antonymes.

La proportion de synonymes que nous avons calculée pour les mots de haute fréquence à n=1 est comparable à celle qu'a obtenue van der Plas (2008) avec ses données, soit 21,31 % (nous obtenons 19,7 %). Sur les mots de fréquences médianes et de basse fréquence, nos résultats divergent davantage : elle obtient en effet 22,97 % et 19,21 %, nous obtenons 14,9 % et 7 %. Les comparaisons pour les autres valeurs de n sont assez peu pertinentes étant donné que chez nous, ce chiffre renvoie aux n premiers voisins alors que chez van der Plas (2008) il renvoie au n-ième voisin.

Pour ce qui est de l'hypo/hyperonymie, nos résultats sont complètement différents. Pour n=1, elle calcule en effet une proportion cumulée d'hyponymes et d'hyperonymes de 32,69 % pour les mots de haute fréquence, de 15,62 % pour les mots de fréquence médiane et de 8,84 % pour les mots de basse fréquence. Le fait que nos scores soient beaucoup plus bas peut probablement s'expliquer par la taille de nos ressources de référence.

4.4.2 La catégorie grammaticale

Dans un deuxième temps, nous avons constitué trois groupes de 1000 noms, verbes et adjectifs sélectionnés de façon aléatoire dans le lexique des VDW (le seul filtre étant le seuil de 10 voisins). Les résultats de leur comparaison avec le DES et JDM ont été rapportés au tableau 4.8.

Le premier constat que l'on peut faire concerne la synonymie, qui est ici aussi la relation la plus repérée dans les voisins. On remarque que, dans l'ensemble, la synonymie est mieux captée par les verbes. La proportion de synonymes parmi les noms et les adjectifs varie assez peu : elle est légèrement

	noms			verbes			adjectifs		
	n=1	n=5	n=10	n=1	n=5	n=10	n=1	n=5	n=10
synonymie (DES)	13,8	9,8	8	17,2	11,8	8,9	14,6	8,7	6,3
antonymie	1,3	0,6	0,5	0,7	0,3	0,2	9.1	3,4	2
hypo/hyperonymie	6,6	4,1	3	0,2	0,1	0,1	1,2	0,6	0,5
méro/holonymie	2,6	1,5	1,3	0	0	0	0	0	0
total	24,3	16	12,8	18,1	12,2	9,2	24,9	12,8	8,8

TAB. 4.8 – Recouvrement entre les voisins des noms, verbes et adjectifs et les relations de JDM (en %).

plus élevée pour les adjectifs à n=1, alors que c'est le contraire à n=5 et n=10. Cela pourrait traduire le fait que les adjectifs ont moins de synonymes, mais qu'ils sont légèrement mieux captés par les voisins.

Il ressort également que c'est parmi les voisins adjectivaux que l'antonymie est le mieux repérée. Les proportions pour les noms et les adjectifs sont environ 9 fois inférieures.

Comme on pouvait s'y attendre, l'hypo/hyperonymie est principalement repérée dans les voisins nominaux. Parmi les quelques hypo/hyperonymes repérés dans les adjectifs, on trouve des couples comme *coréen/asiatique*, *catholique/chrétien* ou *anglais/britannique*. En revanche, les deux couples verbaux identifiés comme des hypo/hyperonymes, à savoir *manger/se nourrir* et *soigner/guérir*, sont plus contestables. Les méro/holonymes présentent un cas de figure assez similaire puisqu'ils ne sont repérés que parmi les voisins nominaux (ce qui était également assez prévisible).

4.4.3 Arguments *vs* prédicats

Nous nous sommes ensuite posé la question de savoir si le statut d'argument ou de prédicat – et, plus largement, la nature de la relation accolée au mot cible – avait une influence sur la nature des voisins extraits. Malheureusement, le fait d'imposer un seuil minimal de 10 voisins aux mots cibles a limité nos possibilités. Nous n'avons en effet pu observer le recouvrement entre les voisins et les ressources de référence que dans les cas où le mot cible est :

- un verbe,
  - porteur de la relation sujet ;
  - porteur de la relation objet ;
- un nom,
  - argument ;

	verbes					
	sujet			objet		
	n=1	n=5	n=10	n=1	n=5	n=10
synonymie (DES)	17,1	11,2	8,9	19,2	13,2	10
antonymie	0,9	0,4	0,3	0,8	0,5	0,4
hypo/hyperonymie	0	0	0	0,2	0,2	0,1
méro/holonymie	0	0	0	0	0	0
total	18	11,6	9,2	20,2	14	10,5

TAB. 4.9 – Influence de la relation portée par le verbe sur la composition des voisins.

	noms								
	argument			modifieur			de		
	n=1	n=5	n=10	n=1	n=5	n=10	n=1	n=5	n=10
synonymie (DES)	15,9	11,1	8,5	17,5	10,6	8	5,9	4,4	3,5
antonymie	1,3	0,9	0,7	0,9	0,6	0,5	0,9	0,6	0,4
hypo/hyperonymie	6,7	4,3	3,3	7	4	2,9	4,3	2,4	1,7
méro/holonymie	2	1,4	1,2	2,6	1,8	1,4	1,2	0,9	0,8
total	25,9	17,7	13,7	28	17	12,8	12,3	8,3	6,4

TAB. 4.10 – Influence de la relation portée par le nom sur la composition des voisins.

- un nom porteur de la relation modifieur ;
- un nom porteur de la relation DE.

Dans les autres cas de figure, le nombre de mots cibles était inférieur à 1000[12]. Les adjectifs n'ont pas été étudiés ici étant donné qu'ils n'apparaissent dans la majorité des cas qu'en position argument.

Les données rapportées au tableau 4.9 montrent que les voisins des verbes qui partagent les mêmes sujets ou les mêmes objets ont des compositions comparables. Les voisins des verbes rapprochés par les objets qu'ils prennent en commun semblent contenir une proportion légèrement plus élevée de synonymes.

Le tableau 4.10, en revanche, fait apparaître un phénomène intéressant : alors que les voisins des noms arguments et de ceux qui portent la relation modifieur captent chacune des relations lexicales dans des proportions relati-

[12]Le nombre de verbes porteurs de la relation sujet était de 934, mais nous avons choisi malgré tout de les intégrer à notre étude pour pouvoir les comparer avec les verbes qui portent la relation objet. Les proportions ont été calculées en conséquence.

	voisins		
	argument	modifieur	de
alimentation	*approvisionnement* (s)	*nourriture* (m)	*avoir besoin* SUJ
cave	*grotte* (s)	*caverne* (s)	*loger* DANS
château	*palais* (h)	*maison* (h)	*seigneur* DE
gain	*débit*	*profit* (s)	*économiser* OBJ
moteur	*machine* (h)	*machine* (h)	*accident* DE
position	*situation* (s)	*situation* (s)	*placer* OBJ
prison	*camp*	*cimetière*	*emprisonner* À
reste	*moitié*	*vestige* (s)	*compter* SUJ
signe	*symbole* (h)	*symbole* (h)	*symboliser* OBJ
symptôme	*lésion*	*affection*	*diagnostic* DE

TAB. 4.11 – Comparaison des meilleurs voisins d'une série de noms portant les relations argument, modifieur et DE.

vement comparables, les voisins des noms porteurs de la relation DE affichent des performances environ deux fois moins élevées. On peut supposer que les prédicats composés d'un nom et de la relation DE sont rapprochés de mots avec lesquels ils entretiennent une – ou *des* – relation(s) qui sort(ent) du cadre de celles qui sont mesurées ici. Nous avons donc cherché à comprendre quelle était la nature de ces relations. Pour cela, nous avons sélectionné dix mots qui apparaissent dans nos trois listes de 1000 noms et nous avons rapporté leurs voisins pour n=1 au tableau 4.11. Nous avons noté s les mots qui ont été captés comme des synonymes du mot-cible, h ceux qui ont été captés comme des hypo/hyperonymes et m les méro/holonymes. De plus, pour une meilleure interprétation des données, nous avons fait apparaître la relation syntaxique des voisins des noms en *de* (ce qui était inutile pour les deux autres types de noms puisqu'ils ne peuvent être rapprochés que de noms porteurs de la même relation).

On peut voir que les noms qui portent la relation DE ont une certaine propension à être rapprochés de verbes. Les rapports entre les noms et ces verbes sont assez hétérogènes. Ils reflètent la polysémie de la préposition *de*, qui peut – entre autres – exprimer :

- la localisation (*cave*_DE et *loger*_DANS partagent des contextes comme *villa*, *hôtel* ou *palais*) ;
- un lien de type SUJET ou OBJET, dans le cas des noms déverbaux. On observe ainsi la formation de variantes verbo-nominales (Fabre, 2010), c'est-à-dire de couples nom/verbe qui partagent le même contenu informationnel. Ils apparaissent, de fait, dans des contextes similaires :

	VDW			VDLM			VDF		
	n=1	n=5	n=10	n=1	n=5	n=10	n=1	n=5	n=10
synonymie (DES)	18,8	13	10	14,1	9,6	7,2	10	6,6	5,1
antonymie	0,7	0,5	0,3	0,7	0,5	0,3	1,6	0,8	0,5
hypo/hyperonymie	0,3	0,2	0,1	0	0,1	0,1	4,6	2,7	2,1
méro/holonymie	0	0	0	0	0	0	1,5	1	0,8
total	19,8	13,7	10,4	14,8	10,2	7,6	17,7	11,1	8,5

TAB. 4.12 – Influence du corpus sur la composition des voisins.

- *position*_DE/*placer*_OBJ : *curseur*, *lentille*, *rotor*, etc. ;
- *alimentation*_DE/*avoir besoin*_SUJ : *animal*, *plante*, *enfant*, etc. ;
- *signe*_DE/*symboliser*_OBJ : *appartenance*, *renouveau*, *richesse*, etc.

Ainsi, bien que ces variantes partagent la plupart de leur sens, elles sont exclues des dictionnaires de synonymes et ne sont donc pas identifiées par le DES.

On peut également voir que parmi les voisins des noms qui portent la relation DE figurent des noms. La raison pour laquelle ces derniers sont assez mal identifiés par les lexiques est qu'ils relèvent d'une relation non classique.

Les trois cas de couples nom/nom qui figurent parmi nos exemples sont les suivants :

- *château*_DE/*seigneur*_DE : *Sablé*, *Mousson*, *Chantilly*, etc. ;
- *moteur*_DE/*accident*_DE : *moto*, *avion*, *voiture*, etc. ;
- *symptôme*_DE/*diagnostic*_DE : *schizophrénie*, *pathologie*, *infection*, etc.

On voit clairement que ces rapprochements se distinguent de ceux qui sont opérés pour les deux autres types de noms, qui relèvent davantage de la similarité. En effet, si on peut dire que *palais* et *maison* sont des entités de la même nature que *château*, ce n'est pas le cas de *château* et *seigneur*. De même, si *moteur* et *machine* partagent une partie de leur sens, c'est beaucoup moins vrai de *moteur* et *accident*, qui sont liés par une relation thématique particulièrement ténue. Les voisins de *symptôme* rapportés ici présentent le même cas de figure.

4.4.4 La nature du corpus

Pour finir, nous avons comparé 1000 mots – toutes catégories, fréquences[13], etc. confondues – extraits aléatoirement du lexique des VDW, des VDLM et des VDF.

[13]Nous rappelons que puisque les mots doivent avoir un minimum de 10 voisins, les mots très rares sont exclus de fait de la liste des mots cibles potentiels.

Les résultats, rapportés au tableau 4.12, peuvent paraître assez contre-intuitifs. En effet, nous avons vu que les VDF ont été calculés à partir d'un corpus de taille beaucoup plus réduite que pour les VDW/VDLM. De plus, alors que les fréquences moyennes des 1000 mots cibles sélectionnés pour les VDW et les VDLM sont respectivement de 3797 et 3750, celle des 1000 mots cibles extraits des VDF est seulement de 370. Sachant cela, on aurait pu s'attendre à ce que la similarité distributionnelle soit moins bien captée et que, par conséquent, les rapprochements générés soient moins bien identifiés par nos ressources. Or, il n'en est rien : du point de vue du pourcentage global de couples identifiés, les VDF se placent devant les VDLM pour toutes les valeurs de n. En revanche, on observe une variation dans la façon dont se répartissent les voisins : alors que les VDW et les VDLM ne captent pratiquement que des couples de synonymes, les VDF observés ici ont une répartition légèrement plus diffuse. On peut en effet voir que le pourcentage de synonymes n'est que de 10 % dans les VDF – contre 18,8 % et 14,1 % dans les VDW et les VDLM –, mais que la ressource capte mieux les trois autres relations observées, en particulier l'hypo/hyperonymie. À ce stade, il est difficile de donner une explication à ce phénomène. On peut toutefois supposer qu'il s'agit là d'une conséquence liée à la différence de taille des corpus : si la similarité distributionnelle est d'autant mieux calculée que la taille des données est importante, alors il est probable que le manque de données favorise – indirectement – le rapprochement de couples qui relèvent d'un degré de similarité moindre (un hyponyme et son hyperonyme sont forcément moins similaires que deux synonymes).

Préambule méthodologique

Ce préambule constitue une présentation des aspects méthodologiques de la démarche d'observation des voisins que nous adoptons dans les études rapportées dans les quatre chapitres qui suivent. Ces études adoptent en effet un point de vue différent de celui qui caractérise l'approche menée dans le chapitre précédent, dans lequel nous avons mesuré le recouvrement entre les bases de voisins et deux lexiques externes (le DES et JDM).

Les chapitres 5 à 8 portent sur les quatre relations sémantiques que nous avons choisi d'observer en mettant en place quatre protocoles distincts. Toutefois, ces études ont en commun une méthode d'analyse des résultats qui consiste à extraire des échantillons de couples porteurs d'une relation donnée – voisins et non voisins – et à comparer les distributions des mots qui les composent afin de mettre au jour des phénomènes qui pourraient expliquer qu'ils aient été captés ou non par l'ADA. Nous développons ici les choix méthodologiques qui fondent cette démarche.

Identifier les relations parmi les voisins Afin d'observer la façon dont se manifeste une relation donnée parmi les couples de voisins, il nous faut disposer au préalable d'un moyen d'identifier les couples porteurs de cette relation parmi les millions que comptent nos bases. Le principe sur lequel repose l'ensemble de notre approche consiste à extraire ces couples en nous appuyant sur :

- une ressource externe (chapitres 5, 7, 8) ;
- des patrons lexico-syntaxiques (chapitre 6).

Cette démarche nous permet ainsi de disposer de plusieurs jeux de couples qui portent une relation donnée[14] distingués selon qu'ils ont été captés par les voisins ou non. Nous avons rapporté au tableau 4.13 quelques couples de synonymes du DES après leur croisement avec les VDW. Le fait de pouvoir distinguer les couples voisins des non-voisins nous permet de mener des études comparatives afin de mettre au jour les phénomènes qui ont une influence sur le repérage des couples. Toutefois, ces données ne sont pas utilisées en

[14]Ou qui, du moins, ont été filtrés par la référence ou la procédure utilisée.

couples	voisins ?
gouverner/administrer	✓
four/échec	✗
poussière/particule	✓
colère/vengeance	✗
psychique/psychologique	✓
voler/piller	✗
héroïque/noble	✗
foudre/orage	✓

TAB. 4.13 – Couples de synonymes du DES croisés avec les VDW.

l'état. Nous avons en effet identifié deux biais susceptibles d'affecter l'analyse. Afin d'atténuer leur influence, nous avons mis en place deux filtres que nous décrivons ci-après.

Premier filtrage : lexique commun Dans le cas où l'étude part d'un croisement entre une base de voisins et un lexique externe, nous avons fait le choix de ne prendre en compte que les couples de mots qui appartiennent au lexique commun aux deux ressources. Par exemple, le couple *souris/rodenticide* apparaît dans JDM, mais n'est pas capté dans les VDW. Or, le mot *rodenticide* n'apparaît dans aucun couple des VDW. Il ne fait pas partie du lexique des VDW. Ce phénomène peut s'expliquer par le fait que :

- le mot est absent du corpus qui a permis de générer la base de voisins (en l'occurrence le corpus Wikipédia) ;
- le mot apparaît dans le corpus mais apparaît dans un nombre de contextes trop faible pour que sa distribution soit rapprochée de celle d'un autre mot par l'AD.

De ce fait, l'analyse de ces cas de figure a un intérêt assez limité. En les mettant de côté, nous avons privilégié l'étude de couples comme *souris/queue*, qui sont constitués de mots qui ont des fréquences élevées dans le corpus mais dont les distributions ne se recouvrent pas (alors qu'on aurait pu s'y attendre étant donné qu'ils partagent une relation sémantique). Dans ce dernier cas, il est en effet possible de s'appuyer sur l'analyse des contextes d'apparition de chacun de ces deux mots pour expliquer les différences dans leurs distributions. Nous avons ainsi cherché à distinguer les phénomènes qui peuvent s'expliquer par la couverture lexicale du corpus de ceux qui relèvent de la distribution des mots.

Second filtrage : le rapport de productivité (*Rprod*) Ce second filtrage vise à atténuer les effets liés à la mesure de similarité utilisée pour rapprocher les voisins. En effet, à la section 2.2.2, nous avons évoqué les travaux de Weeds (2003) qui montrent que les mesures de similarité comme le score de Lin ont tendance à rapprocher davantage les mots dont les fréquences sont comparables (et ce malgré les mesures de pondération qui sont mobilisées lors de ce calcul). De ce fait, une certaine proportion des couples de synonymes/antonymes/hyponymes/méronymes – ceux qui sont constitués de mots dont les fréquences sont déséquilibrées – ont statistiquement moins de chance d'être captés par l'ADA. Cela complique considérablement l'analyse des couples de mots qui n'ont pas été extraits par l'ADA. Il est ainsi problématique d'expliquer le non-repérage par l'ADA de couples comme *réprobation/critique* par le fait qu'ils apparaissent dans des contextes de différentes natures étant donné qu'ils ont des fréquences respectives de 199 et 21 086 dans le corpus Wikipédia. En effet, ces couples ont, d'emblée, peu de chances d'être extraits comme des voisins. Par la suite, nous avons cherché à nous focaliser sur l'analyse des couples dont le non-repérage ne peut pas être imputé à un problème de fréquence mais à un décalage distributionnel révélateur d'un phénomène linguistique. Nous avons alors écarté de nos échantillons les couples les plus *déséquilibrés* : ceux que nous avons voulu étudier en priorité sont ceux qui avaient toutes les chances d'être captés par l'ADA mais qui ne l'ont pas été.

Plutôt qu'en fonction de la fréquence, nous avons choisi de filtrer les couples en fonction de leur productivité. Pour rappel, la productivité désigne le nombre de contextes différents dans lesquels apparaît un mot. Cette valeur nous paraît en effet plus représentative de la distribution d'un mot que sa fréquence d'apparition dans le corpus : en théorie, un mot peut avoir une fréquence très élevée mais une productivité très faible. Dans la pratique, ces deux valeurs sont toutefois très liées (Morlane-Hondère et Fabre, 2012).

Nous avons distingué les couples les plus susceptibles d'être captés par l'ADA de ceux qui sont pénalisés par leur productivité en calculant le rapport entre la productivité – *Rprod* – de leurs deux membres. Par exemple les synonymes *monarque* et *chef* – non voisins – ont des productivités respectives de 59 et de 1941, leur Rprod est donc de 0,03 (soit le résultat de la division de 59 par 1941). Nous nous intéressons davantage à l'analyse de couples non voisins dont les productivités sont moins inégales comme *aspect/air*, qui apparaissent respectivement dans 883 et 935 contextes différents, et dont le Rprod est de 0,94. Dans ce cas, le non-repérage par l'ADA peut s'analyser comme un véritable décalage distributionnel lié à un phénomène linguistique. Nous avons ainsi appliqué aux couples à analyser un seuil de Rprod minimal afin de ne conserver que ceux qui possèdent les propriétés optimales

pour l'ADA. La valeur de ce seuil est différente d'une étude à l'autre (nous expliquons dans chacune des études la façon dont nous l'avons fixée).

Interprétation des résultats La phase d'interprétation consiste à s'appuyer sur la comparaison des contextes d'apparition de deux mots pour expliquer le fait qu'ils aient été captés – ou non – par l'AD. Pour ce faire, nous nous sommes servi des interfaces de consultation des VDW et des VDLM accessibles sur la plate-forme REDAC. La version des VDF intégrée dans l'application Les Voisins D'En Face[15] – qui permet de comparer les VDF et les VDLM – n'étant pas la même que celle sur laquelle nous avons travaillé, nous avons dû consulter la base en local, à l'aide d'un programme Perl de notre conception.

À titre d'illustration, nous avons rapporté aux figures 4.4 et 4.5 les dix contextes qui ont l'information mutuelle la plus élevée avec – respectivement – les arguments *établissement* et *restaurant* tels qu'ils apparaissent dans l'interface de consultation des VDW. La figure 4.6 rapporte un extrait des contextes communs à ces deux mots.

Le fait de pouvoir accéder aux contextes d'apparition des mots nous donne ainsi une image de la façon dont ils s'emploient dans le corpus. Cela nous permet de pouvoir expliquer les raisons qui font que deux mots ont été ou n'ont pas été extraits comme des voisins. Dans les analyses que nous menons dans les chapitres suivants, nous illustrons les décalages observés à l'aide des quelques contextes que nous estimons les plus *parlants* (le nombre de contextes extraits pour chaque mot se comptant le plus souvent en centaines, il serait inenvisageable de les reporter, pour chaque analyse, dans leur intégralité).

Protocoles adoptés dans les chapitres suivants Comme nous l'avons évoqué au début de ce préambule, les quatre études qui suivent portent sur quatre relations sémantiques : la synonymie, l'antonymie, l'hyperonymie et la méronymie. Nous verrons que ces relations se manifestent de façons différentes entre les mots qu'elles relient. De ce fait, nous avons décidé de les étudier en mettant en place des protocoles adaptés à chacune d'elles.

L'ordre dans lequel ces études sont présentées est représentatif de la complexité des procédés mis en œuvre pour accéder aux phénomènes que nous avons analysés :

– dans le chapitre 5, nous menons une simple comparaison des données du DES avec nos trois bases de voisins afin de montrer que les bases distributionnelles peuvent permettre d'*adapter* le dictionnaire à certains

[15] `http://redac/applications/vdef.html`

Prédicat			Argument			
Catégorie	Lemme	Relation	Catégorie	Lemme	IM ⬆ ⬇	Fréquence ⬆ ⬇
V	différencier	de	N	établissement	9.165	5
V	gérer	comme	N	établissement	8.066	8
N	inscription	dans	N	établissement	7.83	5
V	assujettir	obj	N	établissement	7.683	5
N	élève	dans	N	établissement	7.683	5
N	visite	dans	N	établissement	7.589	6
V	dispenser	dans	N	établissement	7.512	9
V	scolariser	dans	N	établissement	7.491	6
N	visibilité	de	N	établissement	7.248	5
V	agréer	obj	N	établissement	7.198	7

FIG. 4.4 – Visualisation des contextes d'apparition de l'argument *établissement*.

Prédicat			Argument			
Catégorie	Lemme	Relation	Catégorie	Lemme	IM ⬆ ⬇	Fréquence ⬆ ⬇
V	manger	dans	N	restaurant	10.243	8
V	dîner	à	N	restaurant	10.243	10
V	dîner	dans	N	restaurant	10.243	8
N	serveur	dans	N	restaurant	9.732	9
N	cuisinier	de	N	restaurant	9.311	13
N	menu	de	N	restaurant	8.538	12
V	étoiler	obj	N	restaurant	8.237	7
V	manger	à	N	restaurant	8.163	8
N	lé	mod	N	restaurant	8.129	18
N	café	mod	N	restaurant	7.559	49

FIG. 4.5 – Visualisation des contextes d'apparition de l'argument *restaurant*.

Catégorie	Lemme	Relation	IM moy ⬆ ⬇
V	fermer	suj	5.51
V	trouver	de	5.23
N	propriétaire	de	5.05
V	fréquenter	obj	4.67
V	travailler	dans	4.65
V	servir	dans	4.48
V	accueillir	suj	4.44
V	ouvrir	suj	4.40
V	abriter	obj	4.37
V	transformer	en	4.35

FIG. 4.6 – Visualisation des contextes communs aux arguments voisins *établissement* et *restaurant*.

types de textes ;
- le chapitre 6 implique l'utilisation de patrons lexico-syntaxiques définis pour capter la relation d'antonymie. Nous croisons les couples extraits par ces patrons avec les voisins afin de tester l'hypothèse d'un double fonctionnement syntagmatique/paradigmatique des antonymes ;
- dans le chapitre 7, nous montrons comment l'ADA peut nous informer sur la catégorisation des mots dans les textes en mesurant la proportion d'hyponymes de JDM extraits pour chaque hyperonyme. Une deuxième mesure nous permet d'étudier la variation de cette proportion entre les trois bases de voisins ;
- la méthodologie adoptée dans le chapitre 8 est la plus complexe des quatre que nous avons mises en place. Nous cherchons en effet à étudier les modalités du repérage des couples de méronymes en nous appuyant sur la nature sémantique des mots qui les composent, ce qui implique de mobiliser une ressource sémantique externe.

Chapitre 5

Utiliser des bases distributionnelles pour filtrer les synonymes du DES

Sommaire

Dans le chapitre 4, nous avons opéré un croisement entre les bases de voisins et le Dictionnaire électronique des synonymes (DES), ce qui nous a permis de mesurer la quantité de couples de voisins qui (ne) sont (pas) des synonymes et, réciproquement, celle de synonymes qui (ne) sont (pas) des voisins. Nous avons ainsi pu mettre en lumière un décalage flagrant entre une ressource comme le DES et les voisins : à lexique partagé, 41,6 % des synonymes du DES étaient captés par les voisins de Wikipédia (VDW), 33,7% par les voisins de Le Monde (VDLM) et 21,1 % par les voisins de Frantext (VDF). Dans tous les cas, la majorité des synonymes ne sont pas repérés. Or, ces couples de synonymes sont, par définition, censés être substituables dans un ensemble plus ou moins étendu de contextes. Quelles sont alors les modalités qui font que certains synonymes sont extraits comme des voisins distributionnels et pas d'autres ?

Nous proposons ici d'aborder cette problématique à travers l'étude des effets du filtrage d'un dictionnaire de synonymes par une ressource distributionnelle. Notre démarche s'appuie sur l'hypothèse selon laquelle une base distributionnelle peut permettre de *sélectionner* les synonymes d'un dictionnaire en présentant à l'utilisateur ceux qui sont les plus immédiatement substituables à un mot donné et pour un type de texte donné. On peut par exemple supposer qu'un dictionnaire de synonymes intégré à un outil d'aide à la rédaction utilisé par des journalistes puisse bénéficier de l'apport d'une base distributionnelle comme les VDLM (qui ont été calculés à partir de textes journalistiques).

Nous présentons plus en détails l'intérêt de cette démarche à la section 5.1 puis nous donnons un aperçu des effets du filtrage sur les données du DES à la section 5.2. Après avoir sélectionné un jeu de synonymes dans le DES – section 5.3 –, nous procédons à une analyse linguistique des effets du filtrage de ces synonymes par les VDW (section 5.4). Nous décrivons ensuite, à la section 5.5, l'impact de la nature du corpus sur le filtrage des synonymes en croisant le DES et les VDLM/VDF. Nous concluons ce chapitre en évoquant, à la section 5.6, la question de l'évaluation du filtrage.

5.1 Intérêts du filtrage

Nous avons jusque-là envisagé le DES comme une ressource lexicale nous permettant d'évaluer des bases distributionnelles. Or, ce lexique, consultable sur le Web, constitue pour de nombreux utilisateurs un outil d'aide à la rédaction. Manguin (2005) rapporte que les administrations québécoises et

Fig. 5.1 – Mode de présentation des synonymes du DES – ici, ceux de l'adjectif *remarquable* (extraits) – sur la plate-forme CNRTL.

suisses sont parmi les plus grands utilisateurs du DES. L'article de Orosemane (2012) témoigne également de son utilisation dans le milieu journalistique.

Lancée en 1998, la version en ligne du DES a connu un succès grandissant, si bien qu'en octobre 2012, le site reçoit 200 000 requêtes par jour (Orosemane, 2012). Le dictionnaire peut être consulté soit *via* le site du CRISCO[1], soit sur la plate-forme CNRTL[2]. Dans les deux cas, l'interface de consultation consiste en un champ unique dans lequel l'utilisateur saisit sa requête. Comme on peut le voir à la figure 5.1, les synonymes du mot donné en requête sont présentés dans une liste ordonnée. Il est à noter que le site du CRISCO ne fournit que la liste ordonnée des dix premiers synonymes, les autres étant classés par ordre alphabétique. L'ordre dans lequel sont présentés les synonymes est défini selon un calcul de proximité basé sur le principe suivant :

> [S]i un synonyme recouvre beaucoup de sens élémentaires du mot-vedette, il est assez proche de ce dernier au point de vue sémantique. (Manguin *et al.*, 2004, p. 6)

Il s'agit donc de s'appuyer sur le réseau que forment les synonymes pour faire apparaître en début de liste ceux qui partagent le plus de cliques (cf. p. 54) avec le mot vedette.

Ce mode de présentation pose un problème majeur : l'ordre dans lequel sont présentés les synonymes est statique, dans le sens où il ne prend pas en compte le contexte dans lequel se trouve le mot dont l'utilisateur cherche un synonyme. Sans cette information, il est impossible de prédire quels synonymes répondront le mieux au besoin de l'utilisateur. Or, on sait que tous les synonymes d'un mot ne sont pas également adaptés à un contexte donné (Murphy, 2003) : par exemple, parmi les synonymes de *retrait*, *repli* conviendrait dans un texte d'actualité qui évoque la progression d'une force armée alors que *reflux* sera plus adapté à un texte relatif au domaine maritime,

[1] `http://www.crisco.unicaen.fr/des/`
[2] `http//www.cnrtl.fr/synonymie/`

de la même façon qu'*abolition* sera plus pertinent dans un texte de loi ou *amputation* dans un texte relatif à la médecine. Le fait de mettre en avant les synonymes les plus probables constitue donc un pis-aller face à l'absence d'informations sur le contexte d'apparition du mot qui a été donné en requête (ou du contexte dans lequel va s'insérer le synonyme).

Nous avons vu que l'analyse distributionnelle automatique (ADA) permettait de synthétiser les contextes d'apparition d'un mot dans un corpus donné et de rapprocher les mots qui partagent ces contextes. Nous proposons donc ici d'utiliser cette méthode pour réorganiser dynamiquement les synonymes proposés par le DES. Il est à noter que le but vers lequel nous tendons est moins ambitieux que celui de l'acquisition : notre approche consiste à combiner le DES avec les voisins, qui jouent alors le rôle de filtre. Ainsi, plutôt que de renvoyer à l'utilisateur les synonymes les plus probables du mot vedette étant donné un réseau de synonymes construit *in abstracto*, il s'agit de renvoyer les synonymes les plus pertinents pour un type de corpus donné.

5.2 Illustration

À titre d'exemple, nous avons rapporté au tableau 5.1 les synonymes du nom *commission*, du verbe *condamner* et de l'adjectif *primitif*. Ces trois listes ont été croisées avec les VDW. Les synonymes qui n'appartenaient pas au lexique commun – *agio*, *ducroire*, *guelte*... pour *commission*, par exemple – n'ont pas été rapportés. Les synonymes qui apparaissent en gras ont été identifiés comme des voisins du mot vedette dans les VDW. Les autres, bien qu'il apparaissent dans le lexique des VDW, n'ont pas été identifiés comme des voisins du mot vedette.

Ces exemples mettent en lumière l'influence du corpus – ici, le corpus Wikipédia –, qui laisse sa *marque* sur la liste de synonymes établie hors corpus. En effet, ces trois mots présentent tous un certain degré de polysémie et l'on peut voir que, dans certains cas, le corpus *désactive* une acception donnée :

- on voit par exemple clairement que *commission* ne partage pas suffisamment de contextes d'apparition avec *prime*, *rémunération*, *gain*, *salaire* ou *paiement* pour qu'ils soient captés par l'ADA. La distribution de ces synonymes, aussi étendue soit-elle, ne recoupe pas celle de *commission*. Or, ces mots relèvent d'une acception de *commission* bien particulière, celle de "rétribution perçue par le commissionnaire"[3]. On peut donc en déduire que dans le corpus Wikipédia, le mot *commission*

[3]Les définitions que nous donnons par la suite sont toutes issues du Trésor de la Langue Française (TLF).

commission
prime
courtage
rémunération
gain
salaire
attribution
remise
intérêt
délégation
comité
charge
traitement
réunion
pouvoir
paiement
mission
mandat
message
titre
groupe
course
change
bureau
brevet
besoin
tribunal

condamner
critiquer
censurer
désapprouver
fermer
reprocher
forcer
punir
obliger
interdire
imposer
frapper
empêcher
défendre
contraindre
charger
barrer
assujettir
sévir
bloquer
perdre
rejeter
reprendre
réduire
occuper
bannir
vouer
sanctionner
nécessiter
murer
exécuter

primitif
élémentaire
premier
simple
naturel
grossier
brut
initial
rudimentaire
rustique
barbare
ancien
primaire
originaire
antique
sommaire
sauvage
primordial
originel
naïf
essentiel
archaïque
ancestral
préalable
original
obscur
aborigène

Tab. 5.1 – Synonymes des mots *commission*, *condamner* et *primitif*. Ceux qui apparaissent en gras ont été captés par l'ADA (corpus Wikipédia).

n'est pas – ou est peu – employé dans des contextes où il renvoie à une valeur pécuniaire. En revanche, il partage les mêmes contextes que :
  - *délégation*, *comité*, *groupe* ou *bureau*, qui renvoient à l'acception "ensemble de personnes officiellement chargées d'une mission à caractère public" ;
  - *charge*, *message*, *mission* ou *mandat*, qui relèvent du sens "charge qu'une personne reçoit de faire quelque chose".
- le cas de *condamner* est un peu moins évident. On peut voir toutefois que parmi ses synonymes non voisins figurent les verbes *fermer*, *barrer*, *bloquer* et *murer*, qui renvoient au sens "interdire l'accès, l'usage de", que ne semble donc pas porter *condamner* dans le corpus Wikipédia.
- concernant l'adjectif *primitif*, on pourrait interpréter le non-repérage de mots comme *simple*, *naturel*, *grossier*, *brut*, *rudimentaire*, *rustique* ou *barbare* comme une négation de l'acception "qui possède des caractéristiques que l'on attribue généralement aux hommes et aux sociétés des temps les plus reculés ou à leurs productions". Celle qui semble prévaloir parmi les synonymes voisins est "qui est à son origine, qui est le plus ancien" (à travers *original*, *archaïque*, *originel*, *ancestral*, etc.).

Ces quelques exemples illustrent l'intérêt qu'il y a à observer dans le détail le résultat du croisement entre les voisins et un lexique externe. Ces exemples nous ont permis d'avoir un aperçu de la façon dont les fonctionnements en corpus influencent le repérage des synonymes. Nous décrivons ci-après le protocole que nous avons mis en place afin d'étudier ce phénomène de façon plus systématique.

5.3 Sélection des données à analyser

Notre démarche consiste ici à observer *à la loupe* le phénomène de décalage entre le DES et les voisins mesuré dans le chapitre précédent en observant la proportion de synonymes extraits par l'ADA pour une série de mots donnés.

Devant l'étendue du phénomène, il nous a fallu mettre en place un protocole de sélection des synonymes à analyser parmi les deux catégories suivantes, issues du croisement du DES et des voisins :

- les synonymes du DES qui ont été captés par les voisins ;
- les synonymes du DES qui n'ont pas été captés par les voisins.

5.3.1 Filtrage sur la productivité

La démarche de filtrer les couples sur la productivité des mots qui les composent vise à atténuer le biais que nous avons évoqué dans le préambule

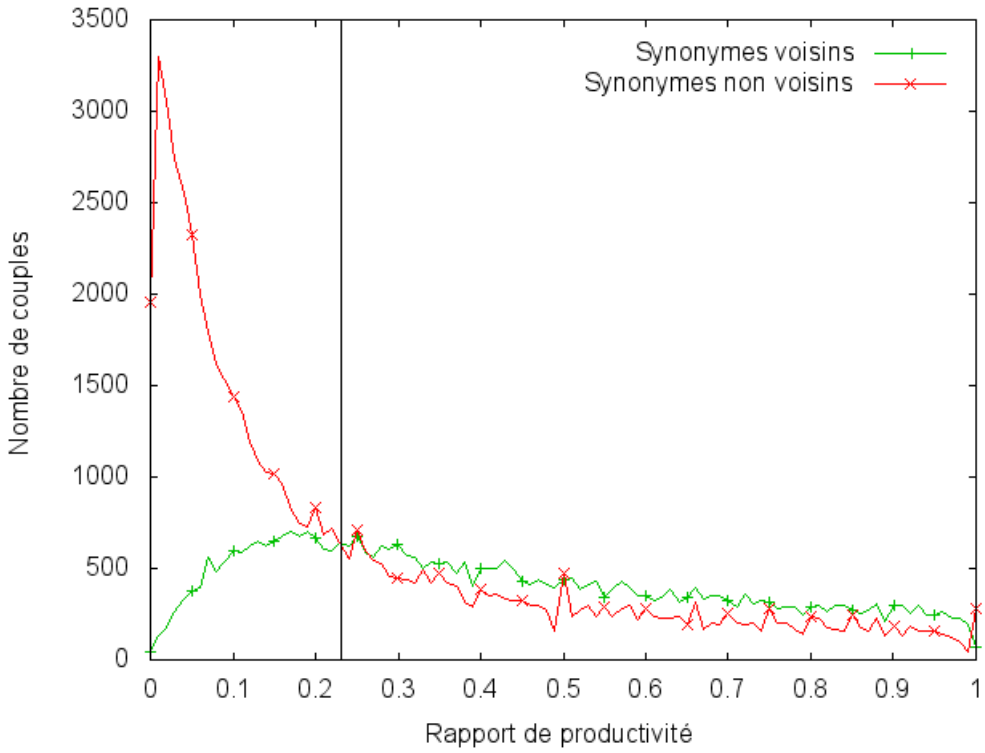


FIG. 5.2 – Évolution du nombre de couples de synonymes voisins et non voisins (VDW) en fonction du rapport de leurs productivités.

méthodologique. Nous avons en effet vu que les mesures de similarité ont tendance à rapprocher les mots qui ont des productivités comparables. Nous cherchons donc ici à écarter de notre échantillon les couples de synonymes dont le rapport de productivté – *Rprod* – est particulièrement déséquilibré. Nous avons défini la valeur du Rprod pour les couples de synonymes en nous appuyant sur la comparaison de la valeur du Rprod pour les couples de voisins et de non-voisins.

Nous avons représenté à la figure 5.2 la façon dont se répartissent les 112 863 couples du DES – dont le lexique est commun à celui des VDW – en fonction de leur Rprod, selon qu'ils sont voisins ou non voisins. On voit clairement que les couples qui ont un Rprod inférieur à 0,23 sont majoritairement des non-voisins. Au delà, le rapport s'inverse. Cela confirme que les couples composés de deux mots dont les productivités sont inégales (et donc dont le Rprod est faible), ont moins de chance d'être captés par l'ADA.

Nous avons donc choisi de filtrer les couples de synonymes en ne conser-

	VDW	VDLM	VDF
noms	1088	1355	837
verbes	574	764	576
adjectifs	263	473	403
total	1925	2592	1816

Tab. 5.2 – Nombre de mots vedettes du DES après filtrage.

vant que ceux dont le Rprod est supérieur à 0,23. Cela concerne :

- 52 092 couples de synonymes dont 58 % sont des VDW ;
- 69 275 couples de synonymes dont 52 % sont des VDLM ;
- 46 189 couples de synonymes dont 20 % sont des VDF.

5.3.2 Filtrage par mot vedette

Une fois écartés les couples de synonymes désavantagés par leur Rprod, notre démarche a consisté à sélectionner une série de mots vedettes dans le DES. Le filtrage effectué à la section précédente a pour effet de diminuer le nombre de synonymes par mot vedette : ce dernier passe de 8 à 5 (dans le cas d'un filtrage par les VDW). De fait, un nombre encore plus important de mots vedettes se retrouvent avec un nombre de synonymes trop bas pour que nous puissions observer des phénomènes comme ceux qui ont été décrits à la section 5.2. Nous avons choisi de ne prendre en compte que les mots vedettes dont le nombre de synonymes est supérieur ou égal à 10 (dans chacune des trois bases de voisins). Le nombre de mots vedettes filtrés par les trois bases de voisins a été rapporté dans le tableau 5.2.

5.4 Effets du filtrage des synonymes

À la section 5.2, nous avons commenté trois exemples qui montraient que certains synonymes d'un mot vedette donné peuvent ne pas partager suffisamment de contextes d'apparition avec ce mot pour qu'ils soient reconnus comme des voisins. Dans cette première étude, nous isolons les mots vedettes dont les synonymes sont le moins bien repérés afin d'identifier les causes de ce phénomène.

Nous avons choisi de nous intéresser aux cas les plus emblématiques du décalage entre le DES et les VDW, à savoir les mots vedettes dont le nombre de synonymes extraits par l'ADA est quasi nul. Nous avons donc calculé la proportion de synonymes identifiés comme voisins par mot vedette et extrait

les 30 noms, verbes et adjectifs dont la proportion est la plus basse, c'est-à-dire inférieure à 0,1 %[4]. Ces derniers constituent notre échantillon d'étude. Ils ont été rapportés au tableau 5.3 (la colonne `syno.` indique le nombre de synonymes du mot vedette). On peut voir qu'on a ici affaire à des cas de décalage particulièrement massifs, l'exemple le plus extrême étant l'adjectif *fou*, dont un seul des 30 synonymes – l'adjectif *malade* – a été extrait comme un voisin distributionnel (ce qui est le cas de la majorité des mots vedettes de notre échantillon, qui ont cependant moins de synonymes).

Cette section est consacrée à la description des différentes causes du décalage entre le DES et les VDW que l'analyse de notre échantillon nous a permis d'identifier. Nous entamons la section par un commentaire de quelques-unes des typologies qui ont été faites des distinctions entre les synonymes (5.4.1). Nous décrivons ensuite chacun des phénomènes entraînant un décalage, à savoir la polysémie (5.4.2), la connotation (5.4.3) et ce que nous avons appelé la *dénotation périphérique* (5.4.4).

5.4.1 Distinguer les synonymes

Cette section constitue un préambule à la description des trois causes de décalage que nous décrivons dans les sections 5.4.2 à 5.4.4. Nous y présentons une série de travaux qui décrivent les différents aspects sur lesquels deux synonymes peuvent diverger. Nous montrons que l'analyse de nos données nous permet d'observer en corpus les conséquences – du point de vue distributionnel – des différences de sens qui distinguent certains couples de synonymes.

D'apparence contradictoire, la question de savoir de quelle façon les synonymes divergent tire sa légitimité du fait que la définition d'une paire de synonymes comme deux mots dont les sens sont parfaitement identiques et qui, de fait, pourraient se substituer l'un à l'autre dans n'importe quel contexte, est purement théorique (dans la mesure où cette substituabilité parfaite serait impossible à démontrer (Cruse, 1986)). De tels synonymes sont dits *parfaits* ou *absolus*. Ainsi, comme le rappelle le passage ci-dessous extrait de la rubrique *Avertissement* de la page de présentation du DES[5], les mots que l'on considère habituellement comme des synonymes ne le sont en fait qu'en partie :

> Suite à des remarques qui nous ont été adressées, il nous semble nécessaire de préciser que deux unités lexicales synonymes ne le

[4]Les mots vedettes dont aucun des synonymes n'était capté ne possédaient, pour la plupart, qu'un nombre trop réduit de voisins pour que l'on puisse faire des analyses pertinentes du fonctionnement de ces mots en corpus.

[5]`http://www.crisco.unicaen.fr/AVERTISSEMENT.html`

noms		verbes		adjectifs	
mot vedette	syno.	mot vedette	syno.	mot vedette	syno.
éclat	29	exciter	29	fou	30
remède	21	vider	22	doux	27
espérance	20	coller	22	rude	27
flamme	20	ôter	20	ferme	25
pli	18	arracher	20	ardent	19
vide	17	ébranler	17	vague	18
ardeur	17	calmer	17	habile	16
clôture	14	altérer	17	brut	16
achèvement	14	allonger	16	honnête	16
attente	14	revêtir	15	épais	15
soumission	14	dérober	15	monstrueux	15
décadence	14	apaiser	14	gentil	15
animation	13	ramasser	14	pénible	15
jet	13	déchirer	13	brusque	14
creux	12	surmonter	12	drôle	14
répétition	12	stimuler	12	douteux	14
lot	12	habiller	12	maigre	13
gage	12	ravir	12	anormal	13
hasard	11	éclaircir	11	abrupt	13
bordure	11	destituer	11	sauvage	13
fable	11	abuser	11	flottant	12
addition	11	aggraver	11	funeste	12
couronnement	11	instruire	11	farouche	12
renvoi	11	tailler	11	sommaire	12
parade	11	chauffer	11	souple	11
coupure	10	prêcher	11	succinct	11
souffle	10	miner	10	vain	11
fraternité	10	aligner	10	vigoureux	10
assainissement	10	bannir	10	stérile	10
complication	10	reculer	10	solennel	10

TAB. 5.3 – Mots vedettes dont les synonymes sont les moins captés dans les VDW.

> sont en pratique que **partiellement**. Cela signifie que c'est **seulement dans un contexte donné** que l'on peut remplacer l'une par l'autre sans modifier notablement le sens de l'énoncé. Cela implique qu'un autre synonyme ne conviendrait pas forcément à ce contexte. Certains mots possèdent ainsi des synonymes à connotation **sexiste**, **homophobe**, **injurieuse**, **raciste** ou qui stigmatisent **l'aspect physique** ou un **handicap**, qui ne peuvent évidemment s'employer dans un contexte neutre, mais qui apparaissent néanmoins ici, puisque ce dictionnaire est censé être le reflet de la langue dans le temps et dans ses divers emplois.

Ainsi, plutôt que de parler de *synonymes*, les lexicologues considèrent plus adéquate l'utilisation des termes *quasi-synonymes* (*near-synonyms ;* Lyons, 1995; Murphy, 2003; Cruse, 2004) ou *parasynonymes* (Lehmann et Martin-Berthet, 2011) pour parler des couples de synonymes qui diffèrent en fonction d'une modalité donnée.

Comme le montrent Inkpen (2003) et Murphy (2003), de nombreux auteurs ont tenté de typologiser la variation qui peut affecter deux quasi-synonymes (Murphy (2003) évoque notamment les travaux de Edmonds (1999), qui a dégagé pas moins de 35 types de variations). Murphy (2003) en distingue six[6] :

- la connotation (*punish/discipline*) ;
- l'affect (*homosexual/gay*) ;
- le registre (*legs/gams*) ;
- le dialecte (*milkshake/frappe*) ;
- le degré de spécialité (*word/lexeme*) ;
- la langue (*dog/perro*).

À titre de comparaison, Cruse (2004) en dégage quatre, de natures différentes :

- l'intensité (*fog/mist*, *laugh/chuckle*, *big/huge*) ;
- pour les verbes, la modalité circonstancielle (*amble/stroll*, *chuckle/giggle*, *drink/quaff*) ;
- l'aspect (*calm/placid*, l'un renvoyant à un état, l'autre à un tempérament) ;
- la différence de "centre prototypique" (*brave/courageous*, le premier renvoyant de façon prototypique à une aptitude physique, le deuxième à une inclinaison morale).

Lehmann et Martin-Berthet (2011) distinguent les différences qui peuvent affecter deux quasi-synonymes selon le niveau d'analyse (ou *plan*) auquel

[6]Les nuances n'étant pas toujours perceptibles dans les traductions françaises des exemples, nous les avons rapportés dans leur version originale.

elles se manifestent :

- les différences de sous-catégorisation et de collocation se situent au niveau **syntaxique**. Par exemple, les modifieurs *grave* et *sérieux* sont substituables quand ils modifient un nom abstrait – *question*, *problème*, *crise* – et non quand ils modifient un nom concret qui porte le sème /audible/ comme *instrument* (contexte dans lequel seul *grave* peut apparaître) ;
- le fait que deux quasi-synonymes ne se distinguent que par rapport à un sème donné relève du plan **sémantique**. Par exemple, *lassitude* et *épuisement* ne se distinguent que par leur degré d'intensité ;
- les distinctions entre quasi-synonymes qui peuvent être à attribuer à une différence de connotation se situent au niveau **pragmatique**. Parmi les exemples que donnent Lehmann et Martin-Berthet (2011, p. 87), on peut citer des couples comme *bru/belle-fille* (variation diachronique), *panosse/serpillière* (variation diatopique), *futal/pantalon* (variation diastratique), *préposé/facteur* (variation langue de spécialité *vs* langue commune) ou *hôtesse de caisse/caissière* (euphémisme).

En permettant de mesurer la similarité entre les contextes d'apparition de deux mots dans un corpus, l'ADA offre la possibilité de vérifier si les nuances de sens qui distinguent les quasi-synonymes se répercutent au point de vue distributionnel. Par la suite, nous nous inspirons des classements évoqués ci-dessus pour organiser la description des trois causes de décalage dans la distribution des synonymes que l'analyse de notre échantillon nous a permis de mettre au jour : la polysémie, la connotation et la dénotation périphérique.

5.4.2 La polysémie

Le phénomène de polysémie correspond à l'explication la plus simple du non-repérage de certains synonymes. En effet, quand deux mots sont synonymes, c'est toujours en fonction d'une acception donnée. Si cette acception ne se manifeste pas – ou peu – dans le corpus, alors les deux mots ne partageront pas suffisamment de contextes pour être extraits par l'ADA. Cette situation peut s'observer sur de nombreux mots de notre échantillon comme *pli*, *clôture* ou *fraternité* :

- *pli* s'emploie dans des contextes comme *oblitérer*_OBJ, *affranchir*_OBJ ou *poster*_OBJ. Le sens qui émerge est celui de "enveloppe renfermant une lettre". Ses premiers voisins sont *levée*, *courrier* et *paquet*. Ses synonymes *arête*, *repli* ou *sillon*, qui renvoient à une acception de *pli* différente, ne sont donc pas extraits par l'ADA ;
- *clôture* s'emploie principalement comme un événement (*jour de*, *gala de*, *cérémonie de*, etc.). Ses contextes d'apparition sont incompatibles

avec des synonymes comme *muraille*, *grille* ou *barrage* ;
- *fraternité* émerge dans le sens de "communauté ou groupement, laïc ou religieux" dans des contextes comme *rejoindre*_OBJ, *membre*_DE ou encore *fondateur*_DE alors que ses synonymes – *charité*, *sympathie*, *confiance*, etc. – réfèrent pour la plupart à son acception "sentiment de solidarité et d'amitié".

Ces observations montrent que la polysémie peut constituer une cause de décalage entre la distribution de deux synonymes. Elles ne permettent toutefois pas de tirer de conclusion plus générale sur les usages que l'on s'attend à observer dans le corpus Wikipédia. En effet, le fait de prendre pour exemple trois mots sémantiquement distincts pour mettre en évidence les conséquences de la polysémie sur le repérage des synonymes ne nous permet pas de mettre le doigt sur des phénomènes plus englobants.

En revanche, l'étude des mots *complication*, *stimuler* ou *anormal* nous a permis d'identifier un phénomène récurrent lié à la présence de textes relevant du domaine médical dans le corpus Wikipédia. Nous décrivons ce cas de figure dans la section 5.4.2.1. Nous évoquons ensuite à la section 5.4.2.2 un autre cas de polysémie cette fois dû au fait que les synonymes renvoient à un sens métaphorique du mot vedette.

5.4.2.1 Les acceptions spécialisées

L'analyse des mots *complication*, *stimuler* ou *anormal* nous a permis d'expliquer la raison du non-repérage de leurs synonymes par le fait que ces trois mots, quand ils sont employés dans le corpus Wikipédia, sont utilisés comme des termes médicaux. Ils apparaissent donc dans des contextes relevant du domaine de la médecine, ce qui bloque le repérage de leurs synonymes, qui réfèrent à des acceptions non spécialisées. Afin d'illustrer le phénomène dont il est question ici, nous décrivons en détail le cas du nom *complication*.

Le nom *complication* apparaît dans les VDW à la fois sous forme d'argument et de prédicat. Nous avons rapporté aux tableaux 5.4 et 5.5 les 10 contextes d'apparition avec lesquels l'argument *complication* et le prédicat *complication*_MOD ont l'information mutuelle la plus élevée ainsi que leurs 10 plus proches voisins (suivant le score de Lin). Des contextes comme *dépister*_OBJ, *mourir*_DE, *prévenir*_OBJ – dans le tableau 5.4 – ou les modifieurs *neurologique*, *cardio-vasculaire* ou *hépatique* – dans le tableau 5.5 – montrent clairement que *complication* adopte ici un sens bien particulier, qui correspond à l'acception "phénomènes pathologiques nouveaux résultant de l'évolution d'une maladie et appelant généralement un traitement particulier". La prépondérance des contextes d'apparition de *complication* relatifs au domaine de la médecine est telle que la grande majorité de ses voisins

contextes d'apparition			
cat.	lemme	rel.	i.m.
V	dépister	OBJ	11,159
N	survenue	DE	10,107
N	prise en charge	DE	8,246
V	survenir	SUJ	7,724
N	suites	DE	7,694
N	prévention	DE	7,217
N	risque	DE	7,94
V	mourir	DE	7,6
V	décéder	DE	6,899
V	prévenir	OBJ	6,793

voisins			
cat.	lemme	rel.	Lin
N	hémorragie	_	0,412
N	infection	_	0,382
N	infarctus	_	0,347
N	brûlure	_	0,334
N	pneumonie	_	0,331
N	lésion	_	0,329
N	insuffisance	_	0,324
N	leucémie	_	0,281
N	paralysie	_	0,274
N	asphyxie	_	0,256

TAB. 5.4 – Contextes d'apparition et voisins du nom *complication* en tant qu'argument. La colonne `i. m.` renvoie à l'information mutuelle calculée entre un mot et son contexte d'apparition.

contextes d'apparition		
cat.	lemme	i.m.
A	rarissime	10,202
A	viscéral	8,913
A	opératoire	8,555
A	infectieux	8,411
A	neurologique	8,306
A	cardio-vasculaire	8,23
A	inutile	7,898
A	pulmonaire	7,853
A	articulaire	7,76
A	hépatique	7,705

voisins			
cat.	lemme	rel.	Lin
N	atteinte	MOD	0,385
N	pathologie	MOD	0,37
N	affection	MOD	0,343
N	défaillance	MOD	0,257
N	anomalie	MOD	0,256
N	lésion	MOD	0,247
N	infection	MOD	0,191
N	trouble	MOD	0,189
N	rival	MOD	0,186
N	risque	MOD	0,18

TAB. 5.5 – Contextes d'apparition et voisins du prédicat *complication*_MOD.

les plus proches relèvent également de ce domaine. Le nom *rival* est une exception, puisqu'il est rapproché de *complication* à travers les modifieurs *redoutable*, *potentiel*, *sérieux*, *possible*, *principal* et *grand*, qui ne relèvent pas particulièrement du lexique médical.

On voit donc aisément pourquoi les synonymes que renvoie le DES pour *complication*, à savoir *piège*, *chaos*, *labyrinthe* ou *combinaison* ne sont pas captés comme des voisins : les contextes dans lesquels *complication* apparaît sont tellement spécifiques que sa substitution par l'un de ses synonymes est inenvisageable. Cela est également vrai pour un synonyme comme *aggravation*, qui semblerait le plus à même de remplacer *complication* dans ce type de contextes (il apparaît dans quelques contextes liés au domaine médical comme *de maladie*, *de état de santé* ou *de traumatisme*).

Nous avons également évoqué le cas de l'adjectif *anormal* et du verbe *stimuler*, qui semblent manifester dans le corpus Wikipédia des sens spécialisés :

- *anormal* porte principalement sur des noms comme *hémoglobine*, *saignement*, *gène* ou *protéine*, qui sont des contextes incompatibles avec des synonymes comme *bizarre*, *paradoxal* ou *insolite* ;
- le cas du verbe *stimuler* est un peu moins emblématique puisqu'un grand nombre de ses contextes ne sont pas spécifiques au domaine de la médecine. Toutefois, on retrouve des termes qui le sont – *production de mélanine*, *sécrétion*, *système immunitaire*, *glande*, etc. – parmi ses objets les plus typiques (ceux avec lesquels il a l'information mutuelle la plus élevée).

Ainsi, alors que cela n'a pas été le cas pour les mots *pli*, *clôture* et *fraternité*, nous avons pu trouver une explication commune au non-repérage des synonymes de *complication*, *anormal* ou *stimuler* par l'ADA. Le fait que le corpus Wikipédia contienne une certaine proportion de textes relevant du domaine médical a une influence sur le sens de ces mots, qui prennent alors un sens spécialisé. En conséquence, la distribution des mots *complication*, *anormal* et *stimuler* est radicalement différente de celles de leurs synonymes dans le DES.

5.4.2.2 Les emplois métaphoriques

Ce cas de figure s'observe sur des mots vedettes comme *flamme* ou *ardeur*. Le cas de *flamme* étant le plus emblématique de ce phénomène, nous le décrivons ci-dessous.

Le nom *flamme* apparaît dans des contextes comme *lécher_*SUJ, *brûler_*AVEC ou *cerner_*SUJ, c'est-à-dire dans son acception "mélange gazeux en combustion, dégageant de la chaleur et généralement de la lumière, pro-

duit par une matière qui brûle". Ici, le décalage naît du fait que la plupart de ses synonymes – *passion*, *désir*, *enthousiasme*, etc. – renvoient à un sens métaphorique de *flamme*. Ces derniers partagent dans le corpus Wikipédia des contextes comme *manifester*_OBJ, *susciter*_OBJ ou *provoquer*_OBJ, dans lesquels *flamme* n'apparaît pas.

5.4.3 La connotation

Nous avons observé dans notre échantillon deux types de connotations que Kerbrat-Orecchioni (1977) désigne sous le nom de "connotation énonciative" et "connotation stylistique" (Le Guern (1972) parle de connotation "sociologique" et "psychologique") . Nous décrivons ci-dessous ces deux cas de figure.

5.4.3.1 Connotation énonciative

Dans ce cas de figure, le mot vedette porte une valeur axiologique que ses synonymes n'ont pas (ou *vice versa*). Ce phénomène est énonciatif dans le sens où l'emploi d'un mot axiologiquement marqué plutôt que d'un mot *neutre* donne des indications "non sur le référent du message, mais sur son énonciateur" (Kerbrat-Orecchioni, 1977, p. 104). Les mots marqués et non marqués partagent le même sens dénotatif. De fait, même si deux mots peuvent diverger par le point de vue que porte le scripteur sur le référent dénoté, ils peuvent tout de même constituer des synonymes potentiels (Murphy, 2003). Il est d'ailleurs intéressant de noter que, déjà à l'époque, Bally (1951, p. 165) reproche aux *dictionnaires idéologiques* le fait qu'"on y rencontre dans un pêle-mêle qui est d'ailleurs un défaut, des expressions de toute nature", c'est-à-dire de façon indifférenciée des expressions "neutres et intellectuelles" comme *fuir*, *s'enfuir*, *s'échapper*, etc. et des expressions "affectives", comme *se sauver*, *filer* ou *prendre ses jambes à son cou*, qui varient par "la manière dont l'idée même de *fuir* est présentée".

Parmi les mots vedettes de notre échantillon, les adjectifs *sommaire* et *fou* nous permettent d'illustrer ce phénomène. Ils ont respectivement pour synonymes *simpliste* et *stupide*, lesquels sont clairement péjoratifs. Ils n'ont pas été extraits par l'ADA. Si l'on regarde de plus près le cas de *fou*/*stupide*, on peut voir que les premiers voisins de *fou* sont *malade*, *immortel* et *endormi* et que ceux de *stupide* sont *intelligent* et *curieux*. Cela confirme le fait que *fou* est utilisé comme un adjectif non marqué alors que *stupide* est rapproché d'adjectifs qui portent une valeur subjective.

Étant donnée la nature du corpus Wikipédia, on peut dans un premier temps s'étonner de la présence même de termes marqués parmi les VDW.

En effet, la rédaction des articles de Wikipédia est soumise au respect de la neutralité de point de vue[7], qui proscrit l'emploi de mots connotés. De fait, ces mots – ici, l'adjectif *stupide* – apparaissent en majorité dans des contextes bien particuliers comme des citations – en discours direct (1) ou rapporté (2) – ou des titres d'œuvres (3). Les cas où ces mots expriment le point de vue du rédacteur – comme c'est le cas dans l'exemple (4) – restent assez marginaux (la phrase rapportée dans cet exemple ne figure d'ailleurs plus dans la version actuelle de l'article concerné[8]).

(1) Lorsque le détective privé tire sur Julian Marty, il lui dit "Qui a l'air **stupide** maintenant ?".

(2) Lors d'une interview le 13 Mars 2005, Augmon s'est également emporté contre un journaliste du Orlando Sentinel, qualifiant sa question de **stupide**.

(3) En 1967, toujours, Sacha Distel, interpréta la version française de la chanson, "Ces mots **stupides**", en duo avec Joanna Shimkus.

(4) Pourtant, les Sardaukars ne sont pas **stupides**, et savent tirer leçon de leurs erreurs, ou arrêter le combat quand ils reconnaissent que tout espoir est perdu.

Le texte rapporté en citation s'affranchit des contraintes liées au genre encyclopédique : il peut contenir des traces de subjectivité, relever de n'importe quel registre. De ce fait, la spécificité de ces contextes peut créer un décalage distributionnel entre les synonymes axiologiquement marqués, qui apparaissent principalement dans des citations, et leurs équivalents non marqués. Ces cas de décalage sont ainsi révélateurs d'un certain degré d'hétérogénéité du corpus Wikipédia.

5.4.3.2 Connotation stylistique

Un des inconvénients du DES est que les synonymes sont fournis sans indication d'usage. Or, le fait qu'une relation de synonymie porte sur deux mots qui relèvent de registres différents a forcément des conséquences sur leurs distributions, et donc, sur leur substituabilité. Ainsi, dans son étude diachronique des dictionnaires de synonymes Ferrara (2010) note que "les registres de langue n'étaient pas jugés comme substituables jusqu'au milieu du XX^{e} siècle lorsqu'il s'agissait de synonymie" (p. 9). De la même façon, Petit (2005) note que si deux mots appartenant à des registres différents

[7] `http://fr.wikipedia.org/wiki/Wikipédia:Neutralité_de_point_de_vue`
[8] `http://fr.wikipedia.org/wiki/Sardaukar`

peuvent présenter des "propriétés sémiotiques" identiques, cela n'implique pas qu'ils seront substituables dans tous les contextes.

Nos données nous permettent de vérifier ce principe. En effet, nous avons pu constater que certains mots vedettes de notre échantillon possèdent des synonymes qui ne sont pas reconnus par l'ADA à cause du fait que ces synonymes renvoient à un registre de langue qui n'est pas employé dans le corpus Wikipédia. Il est à noter que les mots qui relèvent intrinsèquement du registre argotique – *bagnole*, *godasse*, *clamser* – n'apparaissent pas parmi les synonymes des mots vedettes de notre échantillon : leur fréquence dans le corpus est quasi nulle. En revanche, certains mots de la langue *standard* peuvent prendre des emplois argotiques, comme c'est le cas de *veine*, qui figure dans le DES comme un synonyme de *hasard*. Or, *veine* s'emploie dans le corpus dans des contextes comme *couler*_DANS, *injection*_DANS, *occlusion*_DE ou peut être modifié par les adjectifs comme *splénique*, *fémoral* ou *jugulaire*. Ces contextes nous montrent clairement que *veine* est employé dans le corpus comme un type de vaisseau sanguin et non au sens de "chance, fortune". Il n'apparaît pas dans les même contextes que *hasard* comme *part*_DE, *jeu*_DE, *rencontrer*_PAR, etc.

On peut tirer les mêmes conclusions pour le mot vedette *chauffer*, qui, dans le DES, a pour synonymes *voler* et *dérober*, qui renvoient à une acception argotique du mot.

5.4.4 La dénotation périphérique

À la suite de Murphy (2003), nous distinguons dans le sens dénotatif d'un mot son noyau sémantique (*core meaning*) et ses – éventuels – traits périphériques (*peripheral features*). La notion de trait périphérique nous permet en effet d'aborder le problème des synonymes qui partagent un même noyau de sens (ce qui n'était pas le cas dans la section 5.4.2 : un *pli* (postal) n'est pas la même chose qu'un *repli*) mais qui se distinguent du point de vue des nuances sémantiques qu'ils peuvent véhiculer (qui sont donc exprimées par ces traits). Il est à noter que nous sommes encore ici dans le domaine de la dénotation : les nuances connotatives sont abordées à la section 5.4.3. Nous décrivons ici deux cas de figure.

5.4.4.1 Absence/présence de traits périphériques

Murphy (2003, p. 148) illustre la différence entre noyau de sens et traits périphériques en comparant le sens du verbe *punish* et celui de ses synonymes dans le American Heritage Dictionary. Elle montre que le sens de *punish*, qui apparaît comme le moins spécifique, est inclus dans celui de ses synonymes.

Ces derniers augmentent le degré de spécificité de ce sens en y ajoutant des traits périphériques comme “stresses punishment inflicted by an authority in order to control or eliminate unacceptable conduct” pour *discipline*, “implies corporal punishment or a verbal rebuke as a means of effecting improvement in behavior” pour *chastise*, “usually implies the forfeiture of money or of privilege of gain because rules or regulations have been broken” pour *penalize*, etc. En cela, le fonctionnement de ces synonymes se rapproche de celui des hypo/hyperonymes (ce cas sera spécifiquement traité au chapitre 7).

Parmi les mots vedettes de notre échantillon, le verbe *vider* est représentatif de ce phénomène. Les synonymes qui sont listés par le DES pour *vider* apportent des précisions sur l'action exprimée par le verbe :

- *épuiser* : “vider (quelque chose) de son contenu ou de sa substance” ;
- *écoper* : “vider l'eau qui s'accumule au fond d'une embarcation non pontée, à l'aide d'une écope” ;
- *évacuer* : “vider (un pays, un lieu) des personnes qui l'occupent” ;
- *déménager* : “vider (le meuble, la pièce) de tout ce qu'il contient” ;
- *assécher* : “vider de ses ressources”.

Le parallèle avec l'hypo/hyperonymie est ici d'autant plus marquant que les définitions, qui, pour rappel, sont celles du TLF, se présentent sous la forme *genus/differentiae*, c'est-à-dire qu'elles mentionnent l'hyperonyme – *vider* – puis les nuances qui sont propres à l'hyponyme (nous revenons sur cette dichotomie dans le chapitre 7). Les traits véhiculés par *épuiser*, *écoper*, *évacuer*, etc. apportent des précisions soit :

- sur la nature des compléments du verbe (*l'eau*, *un pays*, *le meuble*...). Ces définitions explicitent ainsi des **restrictions sélectionnelles** qui pèsent sur les synonymes de *vider* ;
- sur la manière dont s'effectue l'action (*à l'aide d'une écope*). On a ici affaire à une **modalité circonstancielle** qui participe à la distinction entre *vider* et *écoper*.

On peut supposer que ces nuances se répercutent sur la distribution de ces synonymes. Ainsi, afin d'observer ce phénomène, nous avons rapporté au tableau 5.6 les 10 arguments[9] qui ont l'information mutuelle la plus élevée avec *vider* et ses synonymes *épuiser*, *évacuer*, *déménager* et *assécher* lorsqu'ils portent la fonction OBJ. Nous n'avons pas fait figurer les contextes du verbe *écoper*. La raison en est que le prédicat *écoper*_OBJ n'a aucun voisin dans les VDW. Le verbe *écoper* n'apparaît que dans le prédicat *écoper*_DE, dont les contextes privilégiés sont *amende*, *suspension*, *pénalité*, etc. On voit donc qu'on a clairement affaire ici à un cas de polysémie.

Les contextes recensés dans le tableau 5.6 confirment l'hypothèse que

[9]Sauf pour *assécher*_OBJ, qui n'en compte que 9 au total.

*vider*_OBJ	*épuiser*_OBJ	*évacuer*_OBJ	*déménager*_OBJ	*assécher*_OBJ
chargeur	deux édition	ville	siège social	marécage
poubelle	pioche	eau	franchise	marais
cuve	gisement	habitant	local	étang
récipient	stock	chaleur	colonel	lac
caisse	munition	population	fois	rivière
coffre	réserve	personne	rue	air
cache	recours	partie	siège	zone
sac	carburant	troupe	capitale	terre
querelle	ressource	territoire	mois	partie
cave	combustible	île	centre	

TAB. 5.6 – Différences dans la distribution de *vider* et de quelques-uns de ses synonymes.

nous avons formulée plus haut, à savoir que les traits portés par les synonymes de *vider* ont une influence sur leurs distributions dans le corpus. Ainsi, alors que *vider* prend pour objets des noms de contenants (*récipient*, *caisse*, *coffre*), *épuiser* s'emploie avec des noms de ressources (*carburant*, *combustible*... *ressource*) ou de *lieux* contenant ces ressources (*stock*, *réserve*, *gisement*), *évacuer* avec des noms de lieux (*ville*, *territoire*, *île*) ou d'animés (*habitant*, *population*, *personne*), *déménager* avec des noms de lieux (*local*, *rue*, *capitale*) et *assécher* avec des noms de lieux qui contiennent habituellement de l'eau (*marécage*, *étang*, *lac*). De ce fait, même en prenant en compte l'ensemble des contextes d'apparition de ces mots dans le corpus Wikipédia, le chevauchement entre les contextes d'apparition de *vider* et de ces synonymes reste très faible.

Ainsi, malgré le fait que *vider* et ses synonymes partagent un même noyau de sens, on s'aperçoit qu'ils ne fonctionnent pas sur le mode de la substituabilité mais de la complémentarité : pour exprimer l'idée de *vider* un lieu de ses habitants on emploie le verbe *évacuer*, le fait de *vider* une ressource se dit *épuiser*, etc. Le fait que *vider* ne porte pas de trait périphérique semble bloquer sa substituabilité avec ses synonymes. Il est à noter que, comme l'évoque Murphy (2003, p. 156), ce phénomène peut être envisagé comme un effet de collocation.

Nous concluons cette section en évoquant l'exemple du mot vedette *aggraver*, qui présente un cas inverse à celui de *vider*. En effet, alors que ce verbe et ses synonymes – *intensifier*, *accroître*, *amplifier* – partagent une valeur d'intensification, seul *aggraver* donne une indication sur la nature – dysphorique – de l'objet. De ce fait, il a une forte tendance à prendre pour objets

des noms négativement marqués comme *traumatisme*, *dépression*, *difficulté* ou *pauvreté*, alors que ses synonymes se construisent avec un spectre de noms beaucoup plus large : *effort*, *présence*, *attaque* pour *intensifier*, *compétitivité*, *efficacité crédibilité* pour *accroître*, *vibration*, *crise*, *puissance* pour *amplifier*.

5.4.4.2 Différence de traits périphériques

Plus haut dans cette section, nous avons fait le parallèle entre la synonymie et l'hypo/hyperonymie en montrant que *vider* et ses synonymes pouvaient être envisagés dans une relation hiérarchique dans la mesure où ces synonymes intègrent le sens de *vider* tout en y apportant des nuances. Nous décrivons maintenant le cas du mot vedette *déchirer*, qui présente un cas de figure légèrement différent : contrairement à ce que nous avons vu pour *vider*, *déchirer* possède des traits périphériques. Nous faisons l'hypothèse que la raison pour laquelle *déchirer* n'a pas été rapproché de ses synonymes *tailler*, *scinder*, et *briser* et que leurs traits périphériques sont incompatibles. Ainsi, la relation entre le mot vedette et ses synonymes n'est plus hiérarchique mais horizontale : elle se rapproche alors de la co-hyponymie.

Chacun de ces mots partage un même noyau de sens, à savoir la notion de séparation d'un tout en plusieurs parties. Toutefois, les conditions dans lesquelles s'effectue cette action diffèrent d'un verbe à l'autre : par exemple, *déchirer* implique que l'action s'effectue à mains nues alors que *tailler* implique un instrument. Les conséquences de cette nuance se répercutent sur le plan distributionnel : certaines matières ne peuvent pas être *déchirées*. De ce fait, la plupart des arguments de *tailler*_OBJ – *vigne*, *bois*, *pierre*, etc. – n'apparaissent pas parmi les contextes de *déchirer*_OBJ – *affiche*, *voile*, *page*. On peut tirer la même conclusion de la comparaison avec les arguments de *briser*_OBJ, qui renvoient à des objets manufacturés – *vase*, *épée*, *chaîne*, etc. –, des parties du corps – *crâne*, *nuque*, *jambe*, etc. – ou des notions – *hégémonie*, *monotonie*, *isolement*, etc.

Il est à noter que parmi les arguments de *déchirer*_OBJ figurent également des noms abstraits. Ce sont principalement des noms de divisions administratives comme *pays*, *royaume* ou *Europe*. En cela, il se rapproche – la valeur dysphorique en moins – de *scinder*_OBJ, qui prend comme arguments des mots comme *duché* ou *paroisse*.

5.4.5 Conclusion

Dans cette section, nous avons décrit les multiples causes qui peuvent expliquer le fait que les synonymes d'un mot recensés dans le DES ne soient pas extraits par l'ADA. Nous avons vu que certaines étaient dues à la compo-

	Synonymes voisins	Synonymes voisins partagés avec les VDW
VDLM	11,4	10
VDF	2,5	2,1

TAB. 5.7 – Nombres moyens de synonymes repérés dans les VDLM et les VDF.

sition du corpus, qui va influer sur le fait que certaines acceptions d'un mot polysémique ou que les mots d'un certain registre de langue auront plus ou moins de chances de se manifester dans les textes (et donc dans la ressource distributionnelle). En revanche, d'autres phénomènes sont à attribuer à des fonctionnements linguistiques : nous avons expliqué le non-repérage des synonymes des verbes *vider* et *arracher* par une incompatibilité en termes de traits dénotatifs. Dans ce cas, les résultats montrent que ces mots qui sont censés fonctionner sur le mode de la substituabilité ont tendance à plutôt fonctionner sur celui de l'exclusion mutuelle, ce qui complique leur extraction par l'ADA.

Dans la section suivante, nous mettons en place un protocole de comparaison entre les trois bases de voisins afin :

1. d'étudier l'influence du type de corpus sur le repérage des synonymes des mots polysémiques ;
2. de vérifier l'hypothèse selon laquelle le repérage de certains types de synonymes reste problématique quelle que soit la nature du corpus utilisé.

5.5 Variation du filtrage en fonction du corpus

Nous avons décrit, dans la section précédente, quelques unes des modalités qui régissent la substituabilité des mots vedettes et de leurs synonymes. Dans cette section, nous montrons que ces modalités ne s'appliquent pas de la même façon selon que la base distributionnelle utilisée a été calculée à partir de textes encyclopédiques, journalistiques, ou littéraires. Nous avons donc, dans un premier temps, comparé successivement la proportion de synonymes extraits comme des voisins par les VDW avec celle des VDLM et des VDF.

Les résultats, rapportés au tableau 5.7, montrent qu'en moyenne, les VDLM extraient 11,4 synonymes par mot vedette (sur 16), ce qui correspond exactement au nombre de synonymes captés par les VDW. En revanche, les

VDF n'en extraient que 2,5. Nous attribuons la faiblesse de ce chiffre à la taille du corpus Frantext. On voit également que la plupart des synonymes présents dans les VDLM/VDF le sont également dans les VDW. Cela montre qu'en moyenne, la plupart des synonymes qui sont extraits par les VDLM et les VDF le sont aussi par les VDW. On note que la réciproque n'est pas vraie pour les VDF puisque les VDW reconnaissent en moyenne 4,6 fois plus de synonymes comme des voisins distributionnels que les VDF. Le fait qu'on observe un recouvrement important entre les VDW et les VDLM montre que les fonctionnements des mots vedettes et de leurs synonymes dans les corpus Wikipédia et Le Monde ne sont pas radicalement différents. Toutefois, ces résultats globaux ne doivent pas masquer le fait que les effets du corpus se font malgré tout ressentir de façon importante pour une certaine proportion de mots vedettes. À titre d'exemple, 16 % des 1202 mots vedettes observés possèdent au moins 5 synonymes qui n'ont été captés que par les VDW ou les VDLM. C'est notamment le cas du mot vedette *ton*, que nous décrivons ci-dessous.

Nous avons rapporté dans le tableau 5.8 la liste des synonymes du nom *ton* selon qu'ils sont :

- captés par les bases de voisins (✓) ;
- présents dans leur lexique mais non repérés comme des voisins (✗) ;
- absents de leur lexique (∅).

Ce tableau appelle plusieurs observations :

- on constate que, sur 35 synonymes, beaucoup sont absents des lexiques des trois ressources. Cela est dû à la rigueur du filtrage opéré en amont. Ainsi, le mot *main*, qui apparaît dans la version non filtrée des VDW, est absent de la version seuillée (avec un Rprod à 0,23), qui ne contient que les couples les plus susceptibles d'être captés. Il est à noter que nous n'avons pas fait apparaître les synonymes qui n'apparaissent dans aucun des lexiques des trois ressources ;
- alors que le nombre de synonymes qui sont captés comme des voisins distributionnels de *ton* est de 14 dans les VDW et de 13 dans les VDLM, il n'est que de 4 dans les VDF. Il s'agit d'une conséquence du fait que la taille du corpus Frantext est très inférieure à celle des corpus Wikipédia et Le Monde (pour rappel, seulement 30 millions de mots contre 200 et 262 millions) ;
- même si les listes de synonymes captés par les VDW et les VDLM sont de tailles à peu près identiques, on constate que leur contenu est relativement différent :
  - moins de la moitié des synonymes captés respectivement par les VDW et les VDLM sont communs aux deux ressources. Ces synonymes communs sont les suivants : *accent*, *écriture*, *goût*, *manière*, *son*. Ces

		VDW	VDLM	VDF
	accent	✓	✓	✓
	air		✓	✓
	bruit	✗	✗	✗
	corde	✗		✗
	couleur		✓	✗
	écho	✗	✓	
	écriture	✓	✓	
	expression		✓	✓
	façon	✓	✓	✗
	facture	✗	✗	
	forme			✗
	genre		✓	✗
	goût	✓	✓	✗
	griffe	✗		
	main		✗	✗
	manière	✓	✓	✗
	mode		✗	
	musique		✗	✗
	note	✓	✗	✗
	nuance	✓	✗	
	parole	✓	✗	✗
	patte	✓		✗
	plume	✓	✗	✗
	procédé	✗	✗	
	signature	✓	✗	
	son	✓	✓	✗
	style		✓	
	teinte	✓		
	tension	✗	✗	
	timbre	✗	✗	
	tonalité	✓		
	touche	✗	✗	
	tour		✗	✗
	verbe	✗		
	voix		✓	✓
Synthèse	✓	14	13	4
	✗	10	15	16
	∅	11	7	15

TAB. 5.8 – Synonymes du nom *ton* selon qu'ils sont captés par les bases de voisins (✓), qu'ils sont présents dans leur lexique mais non repérés comme des voisins (✗) ou qu'ils sont absents de leur lexique (∅).

mots partagent donc les mêmes contextes d'apparition que *ton* que ce soit dans le corpus Wikipédia ou dans le corpus Le Monde ;
- à l'inverse, les mots *nuance*, *signature*, *parole*, *plume* et *note* ne partagent les mêmes contextes que *ton* que dans le corpus Wikipédia. On peut supposer que ces rapprochements sont à attribuer à une plus grande présence du vocabulaire des arts dans les VDW. Une comparaison des modifieurs les plus fréquents de *ton* dans les corpus Wikipédia et Le Monde va dans le sens de cette intuition : alors que *ton* est fréquemment modifié par des adjectifs comme *majeur*, *mineur*, *chaud*, *bleu* ou *clair* dans le corpus Wikipédia, on trouve parmi ses modifieurs privilégiés dans le corpus Le Monde des adjectifs comme *grave*, *ferme*, *vif*, *modéré* ou *solennel*. Dans le deuxième cas, l'emploi de *ton* privilégié semble donc être le suivant : "qualité de la voix (hauteur, timbre, intensité)".

Le troisième point est le plus important pour notre problématique. Il met en lumière l'importance – parfois négligée – de la nature des corpus qui servent de matériau pour la génération de ressources distributionnelles. On a ici pu déduire de l'analyse des synonymes de *ton* captés par les différentes bases de voisins que le sens dans lequel était employé ce mot différait en fonction de la base distributionnelle considérée. Cette observation traduit une différence d'usage dans les corpus qui ont permis de générer ces ressources distributionnelles.

Afin de pouvoir observer à plus grande échelle ce type de phénomènes sur les mots de notre échantillon, nous avons restreint la liste des mots vedettes à ceux qui apparaissent dans les VDW, les VDLM et les VDF. La liste obtenue compte 1202 mots répartis en 621 noms, 389 verbes et 192 adjectifs. Dans un premier temps, à la section 5.5.1, nous avons extrait un nom, un adjectif et un verbe parmi les mots vedettes qui présentent le plus haut degré de variation et nous avons comparé, pour chacun d'eux, la nature des synonymes extraits comme des voisins dans les VDW, VDLM et VDF. Nous avons ensuite, à la section 5.5.2, proposé un mode de comparaison des effets du corpus sur le filtrage qui consiste à prendre en compte l'ordre dans lequel le score de Lin classe les synonymes captés par les voisins.

5.5.1 Impact du corpus sur la présence/absence des synonymes

Dans la section précédente, nous avons pu voir que certains synonymes n'étaient pas identifiés comme des voisins parce qu'ils référaient à une acception du mot vedette qui n'était pas exprimée dans les VDW. Nous cherchons

Comparaison avec les VDLM					Comparaison avec les VDF				
lemme	cat.	vois. VDW	vois. VDLM	diff.	lemme	cat.	vois. VDW	vois. VDF	diff.
suite	N	28	15	13	moyen	N	34	5	29
faculté	N	19	7	12	action	N	35	6	29
trait	N	25	15	10	charge	N	27	1	26
couronne	N	10	1	9	caractère	N	25	2	23
chasser	V	13	4	9	vue	N	24	2	22
exact	A	16	7	9	position	N	25	3	22
enlever	V	21	12	9	note	N	25	3	22
possession	N	9	1	8	disposer	V	22	1	21
hauteur	N	11	3	8	exposer	V	24	3	21
correspondance	N	11	3	8	principe	N	25	4	21
lever	V	6	14	-8	éclat	N	0	4	-4
presser	V	0	9	-9	goutte	N	0	4	-4
surprenant	A	3	12	-9	apercevoir	V	1	5	-4
drôle	A	1	11	-10	misère	N	3	7	-4
cruel	A	5	15	-10	malheur	N	6	10	-4
éclat	N	1	12	-11	angoisse	N	4	8	-4
peser	V	1	12	-11	attente	N	0	5	-5
assurance	N	3	15	-12	charmant	A	3	9	-6
ennui	N	5	17	-12	tendre	A	0	7	-7
espérance	N	1	14	-13	doux	A	2	9	-7

TAB. 5.9 – Mots vedettes pour lesquels le repérage des synonymes varie le plus entre les bases de voisins (en faveur des VDW dans la partie haute du tableau, en faveur des VDLM/VDF dans la partie basse).

maintenant à mettre en lumière ces effets de corpus par la comparaison des VDW avec deux autres bases distributionnelles respectivement calculées à partir de corpus de natures journalistique (les VDLM) et littéraire (les VDF).

Nous avons rapporté au tableau 5.9 les résultats de la comparaison successive des VDW avec les VDLM et les VDF :

- les mots vedettes dont les synonymes sont davantage captés par les VDW que par les VDLM/VDF apparaissent dans la partie haute du tableau. Ce sont majoritairement des noms renvoyant à des concepts abstraits (*faculté*, *action*, *principe*, etc.) ;
- les mots vedettes dont les synonymes sont davantage captés par les VDLM/VDF que par les VDW apparaissent dans la partie basse du tableau. Les adjectifs et les verbes sont un peu plus représentés ici. On remarque que beaucoup de ces mots vedettes ont en commun le fait de porter une valeur axiologique :
  - *drôle*, *cruel*, *ennui* et *surprenant* parmi les mots vedettes dont les synonymes sont le mieux captés par les VDLM ;

– *charmant*, *doux*, *angoisse*, *malheur* et *misère* parmi les mots vedettes dont les synonymes sont le mieux captés par les VDF.

Cette observation va dans le sens des remarques sur la neutralité des textes du corpus Wikipédia que nous avons faites à la section 5.4.3.1.

Afin de faire apparaître les contrastes les plus manifestes entre les bases, nous avons soustrait, pour chaque mot vedette de notre échantillon, le nombre de ses synonymes captés par les VDLM/VDF au nombre de ses synonymes captés par les VDW. Ce chiffre apparaît dans la colonne `diff.` du tableau 5.9 : s'il est positif, alors les synonymes sont mieux extraits par les VDW que par les VDLM/VDF, et l'inverse si le chiffre est négatif.

Dans les sections suivantes, nous comparons les proportions dans lesquelles ont été captés les synonymes de trois des mots vedettes extraits du tableau 5.9 dans les trois bases distributionnelles, à savoir l'adjectif *doux* (5.5.1.1), le nom *éclat* (5.5.1.2) et le verbe *chasser* (5.5.1.3).

5.5.1.1 Exemple 1 : *doux*

On peut voir au tableau 5.10 que la proportion ainsi que la nature des synonymes de *doux* captés dans les trois bases de voisins varient fortement. En effet, alors que 12 synonymes ont été captés dans les VDF, seulement 3 l'ont été dans les VDW et les VDLM. Si l'on prend en compte le nombre de synonymes présents dans le lexique des voisins mais non extraits comme tels (les croix rouges dans le tableau), on se rend compte que plus de la moitié des synonymes qui apparaissent dans le lexique des VDF ont été identifiés comme des voisins alors que ce n'est le cas que d'environ 11 % des synonymes qui apparaissent dans le lexique des VDW et des VDLM. Cela démontre une tendance claire à la substituabilité entre *doux* et ses synonymes dans le corpus Frantext qui semble beaucoup moins se manifester dans les VDW et les VDLM.

Cela peut s'expliquer par le fait que, dans le corpus Wikipédia, *doux* modifie deux types de noms :

– des noms d'aliments ou d'épices (*paprika*, *fenouil*, *beurre*) ;
– des noms relatifs au climat (*microclimat*, *hiver*, *vent*).

Or, ce type de mots n'a que peu de chance d'être modifié par la quasi-totalité des synonymes de *doux*. Les premiers voisins de *doux* dans les VDW sont *sec*, *frais*, *dur*, *humide* et *chaud*.

La nature des mots que modifie *doux* dans les VDLM est plus hétérogène. On retrouve les deux classes identifiées plus haut mais également des noms qui renvoient à des concepts abstraits comme *torpeur*, *revanche* ou *quiétude*. Les premiers voisins de *doux* dans les VDLM sont *tendre*, *étrange*, *chaud*, *sombre* et *délicat*.

		VDW	VDLM	VDF
	agréable	✓	✗	
	amoureux	✗	✗	✗
	beau	✓		✓
	bon			✗
	calme	✓	✗	✓
	charmant	✗	✗	✓
	cher	✗	✗	✗
	clair	✗	✓	✓
	confortable	✗	✗	
	délicat	✗	✓	✗
	facile	✗	✗	✗
	faible		✗	✓
	fin	✗	✗	✗
	gentil	✗		
	harmonieux	✗		
	heureux	✗	✗	✓
	humain		✗	✓
	joli	✗	✗	✓
	léger	✗	✗	✗
	lisse	✗		
	mou	✗		✓
	musical		✗	
	pacifique	✗	✗	
	paisible	✗	✗	✗
	pâle	✗		✗
	patient	✗	✗	
	sage	✗	✗	
	savoureux		✗	
	serein		✗	
	souple	✗	✗	
	sourd	✗	✗	✗
	tendre	✗	✓	✓
	tiède			✓
	tranquille	✗	✗	✓
Synthèse	✓	3	3	12
	✗	24	23	10
	∅	7	8	12

TAB. 5.10 – Synonymes de l'adjectif *doux* selon qu'ils sont captés par les bases de voisins (✓), qu'ils sont présents dans leur lexique mais non repérés comme des voisins (✗) ou qu'ils sont absents de leur lexique (∅).

Dans les VDF, les noms modifiés par *doux* sont quasi-exclusivement des concepts abstraits qui renvoient par exemple à des états (*hilarité*, *hébétude*, *sérénité*) ou des phénomènes (*frôlement*, *chuintement*, *rayonnement*). On note que les noms d'aliments apparaissent également, mais de façon assez anecdotique. Les premiers voisins de *doux* dans les VDF sont *tendre*, *chaud*, *triste*, *étrange* et *fort*.

On voit ainsi que la nature des noms modifiés par *doux* diverge entre les corpus (surtout entre les VDW et les VDF). Cela implique des rapprochements distributionnels différents : même si des mots comme *chaud* ou *tendre* sont rapprochés de *doux* dans plusieurs bases distributionnelles, on voit que chacune possède ses particularités. On note que dans les VDF, *doux* compte parmi ses cinq premiers voisins *triste* et *étrange*. Cela montre bien que la nature des mots que modifie *doux* dans le corpus Frantext favorise le repérage des adjectifs axiologiquement marqués.

5.5.1.2 Exemple 2 : *éclat*

Nous avons rapporté au tableau 5.11 les synonymes du mot vedette *éclat*. Ces derniers sont le mieux captés par les VDLM.

Éclat est un mot polysémique. Or, on peut voir à travers le sens des synonymes qui ont été captés dans chacune des ressources que ses diverses acceptions ne se manifestent pas de façon équivalente dans les corpus Wikipédia, Le Monde et Frantext.

On remarque par exemple que les synonymes qui ont été extraits par l'analyse de Frantext renvoient principalement au sens de "intensité lumineuse d'une source". Ces synonymes sont *clarté*, *feu*, *flamme*, *lueur* et *lumière*. Ils partagent avec *éclat* des contextes d'apparition comme *projeter*_OBJ, *jaillir*_SUJ, *luire*_SUJ et peuvent être modifiés par des adjectifs comme *aveuglant*, *fiévreux* ou *vif*.

En revanche, au vu de la nature des synonymes captés dans les VDLM, on devine que la distribution de *éclat* dans le corpus Le Monde est plus hétérogène que dans le corpus Frantext. En effet, non seulement les VDLM captent des synonymes de *éclat* au sens de "intensité lumineuse d'une source" (les mêmes que ceux qui ont été repérés dans les VDF), mais ils permettent également de faire émerger les emplois suivants :

- "caractère glorieux d'une action ; mérite dans le comportement" à travers les synonymes *célébrité*, *gloire*, *popularité*, qui partagent avec *éclat* des modifieurs comme *soudain*, *médiatique* ou *certain* ;
- "fragment violemment détaché d'un corps qui explose ou que l'on brise" à travers les synonymes *morceau*, *fragment*, modifiés par *moindre*, *grand* ou *petit* ;

	VDW	VDLM	VDF
animation	✗	✗	
appareil			✗
apparence	✗	✗	✗
beauté	✗	✗	✗
bruit	✗	✓	✓
célébrité	✗	✓	
clarté	✗	✓	✓
coloration	✗		
couleur			✓
débris	✗	✗	
déchet	✗	✗	
distinction	✗	✗	
éclair	✗		✓
effet			✗
épanouissement	✗	✗	
explosion	✗	✗	
extérieur		✗	
fanfare	✗		
feu		✓	✓
flamme	✗	✓	✓
fleur		✓	✓
fracas			✗
fragment	✗	✓	

fraîcheur		✓	✗
gloire	✗	✓	✗
grandeur	✗	✗	✗
honneur		✗	✗
illustration	✗	✓	
lueur		✓	✓
lumière			✓
luminosité	✗		
luxe		✓	
magnitude	✗		
majesté	✗		
morceau		✓	✗
partie			✗
pièce			✗
pompe	✗	✓	
popularité	✗	✓	
prestige	✓	✗	
rayonnement	✗	✗	
relief	✗	✓	
richesse		✗	
rumeur	✗	✓	✗
scandale	✗	✓	
son		✗	✗
splendeur		✗	

✓	1	18	9
✗	28	16	14
∅	18	13	24

TAB. 5.11 – Synonymes du nom *éclat* selon qu'ils sont captés par les bases de voisins (✓), qu'ils sont présents dans leur lexique mais non repérés comme des voisins (✗) ou qu'ils sont absents de leur lexique (∅).

– "(au figuré) caractère brillant, fastueux, luxueux d'une chose, d'une situation" à travers le synonyme *luxe* (avec lequel *éclat* partage des modifieurs comme *discret*, *rare* ou *certain*) ;
– "scandale, querelle" à travers le synonyme *scandale*, modifié par *public*, *médiatique* ou *immense*.

L'analyse des contextes qui ont permis de rapprocher ces synonymes du mot vedette *éclat* nous incite toutefois à nuancer l'idée selon laquelle le fait qu'un synonyme soit capté indique que le mot vedette s'emploie dans le sens de ce synonyme dans le corpus. Par exemple, si l'on prend l'ensemble des modifieurs qu'ont en commun *éclat* et *luxe* – *discret*, *rare*, *grand*, *certain*, *dernier*, *petit*, *nouveau* –, on s'aperçoit que la plupart d'entre eux ont un sens très générique. Un *grand éclat*, un *nouvel éclat*, un *éclat rare*, etc. ne renvoient pas forcément à une interprétation de *éclat* au sens de "caractère brillant, fastueux, luxueux d'une chose, d'une situation" (contrairement à des modifieurs comme *médiatique* ou *public*, qui sont moins ambigus).

Dans le même ordre d'idée, nous décrivons, pour finir, le cas de *pompe*. Le nom *pompe* a comme sens "déploiement de faste, de décorum" et est recensé dans le DES comme un synonyme de *éclat*. Le couple *éclat*/*pompe* figure dans les VDLM. Toutefois, l'analyse des contextes d'apparition de *pompe* dans le corpus Le Monde nous montre que ce rapprochement est dû à un phénomène de polysémie. Comme nous l'avons vu plus haut, la polysémie a plutôt tendance à bloquer le repérage des synonymes. On observe ici le cas contraire puisque *éclat* et *pompe* ont été rapprochés *via* les modifieurs *solaire*, *électrique*, *grand*, *petit* et *nouveau*. Dans ce cas, *pompe* renvoie à un dispositif.

Il aurait en effet été surprenant de voir *éclat* et *pompe* au sens de "déploiement de faste, de décorum" partager les mêmes contextes d'apparition dans nos corpus sachant que l'emploi de *pompe* dans cette acception apparaît comme assez peu répandue à l'heure actuelle (si ce n'est dans l'expression *en grand pompe*). Cet exemple nous donne ainsi l'occasion d'illustrer un certain décalage diachronique entre l'emploi des mots qui constituent certaines des paires de synonymes du DES et leur emploi dans nos corpus.

5.5.1.3 Exemple 3 : *chasser*

On peut voir dans le tableau 5.12 que c'est dans les VDW que l'on trouve le plus grand nombre de cas où le verbe *chasser* est voisin de ses synonymes. L'analyse des contextes d'apparition des synonymes qui sont des VDW montre que ces synonymes ont été captés selon deux modalités que nous décrivons ci-dessous.

Dans un premier cas, les synonymes ont été rapprochés du prédicat *chasser*_DE. Ce dernier se construit systématiquement avec, comme arguments,

	VDW	VDLM	VDF
arracher	✗	✗	✗
balancer			✗
balayer		✗	✗
bousculer		✗	
courir	✗	✗	✗
déménager	✗		
déporter	✓		
déposer	✓	✗	✗
détruire	✓	✓	✗
disperser	✗	✓	
dissiper		✗	✗
écarter	✓	✗	✓
éliminer	✗	✗	
éloigner	✓	✓	✗
envoyer			✗
exclure	✓	✗	
exiler	✓		
expulser	✓	✓	

forcer	✓	✗	✗
glisser		✗	✗
lancer	✓		✗
licencier		✗	
ôter			✗
poursuivre	✓		✗
pousser	✓	✗	
rabattre			✗
reconduire		✗	
rejeter	✓	✗	✗
relancer	✗	✗	
remercier		✗	✗
renvoyer	✓	✗	✗
repousser	✓	✗	✗
se séparer		✗	
séparer	✓	✗	✗
vaincre	✓		
traquer		✗	
vider		✓	✗
✓	17	5	1
✗	6	22	21
∅	14	10	15

TAB. 5.12 – Synonymes du verbe *chasser* selon qu'ils sont captés par les bases de voisins (✓), qu'ils sont présents dans leur lexique mais non repérés comme des voisins (✗) ou qu'ils sont absents de leur lexique (∅).

des toponymes ou des noms communs comme *maison*, *trône* ou *patrie*. C'est également le cas de ses synonymes *poursuivre*_EN, *renvoyer*_EN, *expulser*_DE et *séparer*_DE, qui sont donc extraits comme des voisins de *chasser*_DE. Le synonyme *exclure*_OBJ, lui aussi voisin de *chasser*_DE, est un cas à part : les noms de lieux qu'il prend en objet – *Afrique*, *pays* ou *France* – sont dans ce cas à interpréter comme des acteurs politiques et non comme des lieux. La détection du synonyme *exclure* s'appuie donc dans ce cas sur un phénomène d'ambiguïté.

Dans un deuxième cas, les synonymes ont été rapprochés des prédicats *chasser*_SUJ ou *chasser*_OBJ. Nous avons choisi de traiter ces deux cas conjointement étant donné que les ensembles de noms qu'ils prennent en argument présentent un degré de recouvrement remarquable. Cela montre un degré d'interchangeabilité élevé entre les agents et les patients du verbe *chasser* : dans le corpus Wikipédia, ceux qui chassent sont les mêmes que ceux qui sont chassés. Ces agents et patients sont principalement des noms de peuples (*Hollandais*, *Mongol*, *Turc*), d'individus (*roi*, *prince*, *duc*), de groupes d'individus (*troupe*, *armée*, *population*) ou encore d'animaux (*lion*, *loup*). Le fait qu'un verbe puisse prendre les mêmes noms en position sujet et objet entraîne des rapprochements distributionnels à plusieurs niveaux :

- deux prédicats qui partagent le même lemme peuvent être captés comme des voisins (comme c'est le cas de *chasser*_SUJ et *chasser*_OBJ) ;
- un prédicat peut être rapproché de plusieurs prédicats partageant le même lemme (par exemple, *expulser*_SUJ et *expulser*_OBJ). Ce phénomène s'observe sur plusieurs synonymes de *chasser*. Nous avons rapporté dans le tableau 5.13 les prédicats synonymes de *chasser* et marqué d'une coche les relations de voisinage. On peut ainsi voir que si *chasser* et *écarter* ont été rapprochés, c'est sur la base des objets qu'ils prennent en commun. En revanche, les sujets de *détruire* sont à la fois les sujets et les objets de *chasser* (les objets de *détruire* sont des non-animés comme *récolte*, *artéfact* ou *toiture* qui n'apparaissent ni comme sujets ni comme objets de *chasser*). Des verbes comme *repousser* ou *vaincre* montrent même que deux verbes peuvent être rapprochés à la fois par les noms qu'ils prennent comme objets et par ceux qu'ils prennent comme sujets. Par exemple, dans ce cas, les sujets et les objets que prend *chasser* sont également des sujets et des objets de *repousser*. De ce fait, on peut se poser la question de savoir si cela fait de *repousser* et *vaincre* de *meilleurs* synonymes de *chasser* que *écarter*, avec lequel *chasser* ne partage que ses objets.

La raison pour laquelle le repérage des synonymes de *chasser* est moins efficace dans les VDLM et les VDF que dans les VDW vient du fait que le corpus Wikipédia contient des textes qui relatent des événements histo-

		détruire	*écarter*	*expulser*		*forcer*		*repousser*		*vaincre*	
		SUJ	OBJ	SUJ	OBJ	SUJ	OBJ	SUJ	OBJ	SUJ	OBJ
chasser	SUJ	✓		✓	✓	✓	✓	✓	✓	✓	✓
	OBJ	✓	✓		✓		✓	✓	✓	✓	✓

TAB. 5.13 – Relations de voisinage entre les prédicats *chasser*_SUJ et *chasser*_OBJ et les prédicats synonymes de *chasser*.

riques comme les conflits entre les peuples. De ce fait, *chasser* ainsi que ses synonymes ont tendance à partager les mêmes faisceaux de contextes. Par exemple, dans le corpus Wikipédia, *repousser* partage avec *chasser* des objets comme *envahisseur*, *Maure* ou *Prussien* alors qu'il apparaît plutôt avec des objets comme *limite*, *échéance* ou *date* dans Le Monde et *avance*, *assiette* ou *battant* dans Frantext. Or, ces derniers noms n'apparaissent pas en position argument de *chasser*_OBJ.

5.5.2 Utiliser le score de proximité distributionnelle pour classer les synonymes

Nous avons montré, dans l'introduction de cette section, que deux bases de voisins comme les VDW et les VDLM présentent une tendance globale à extraire les mêmes synonymes pour un mot donné. Ce constat aurait tendance à minimiser l'impact du corpus sur les bases distributionnelles produites. Toutefois, observer les voisins sur la base d'une dichotomie absence/présence revient à ignorer les différences distributionnelles qui peuvent être mesurées à un niveau plus fin par le score de Lin.

En effet, dans le cas où deux bases de voisins captent les mêmes synonymes d'un mot donné, il n'est pas évident que ces synonymes aient un score de similarité distributionnelle avec le mot vedette identique dans les deux ressources. À titre d'exemple, sur les 134 synonymes que propose le DES pour le nom polysémique *tour*, les VDW et les VDLM en ont respectivement capté 23 et 29. Parmi les 23 synonymes extraits par les VDW, 22 ont été extraits par les VDLM. Autrement dit, en apparence, on pourrait croire que le type de base distributionnelle utilisé n'a eu que peu d'impact sur le filtrage des synonymes. Or, la comparaison des scores de similarité qui ont été mesurés entre *tour* et ses synonymes dans les VDW et les VDLM nous montre le contraire.

Nous avons comparé à la figure 5.3 :

– l'ordre dans lequel sont proposés les 30 premiers synonymes de *tour* par le DES (dans sa version intégrée à la plate-forme CNRTL) ;

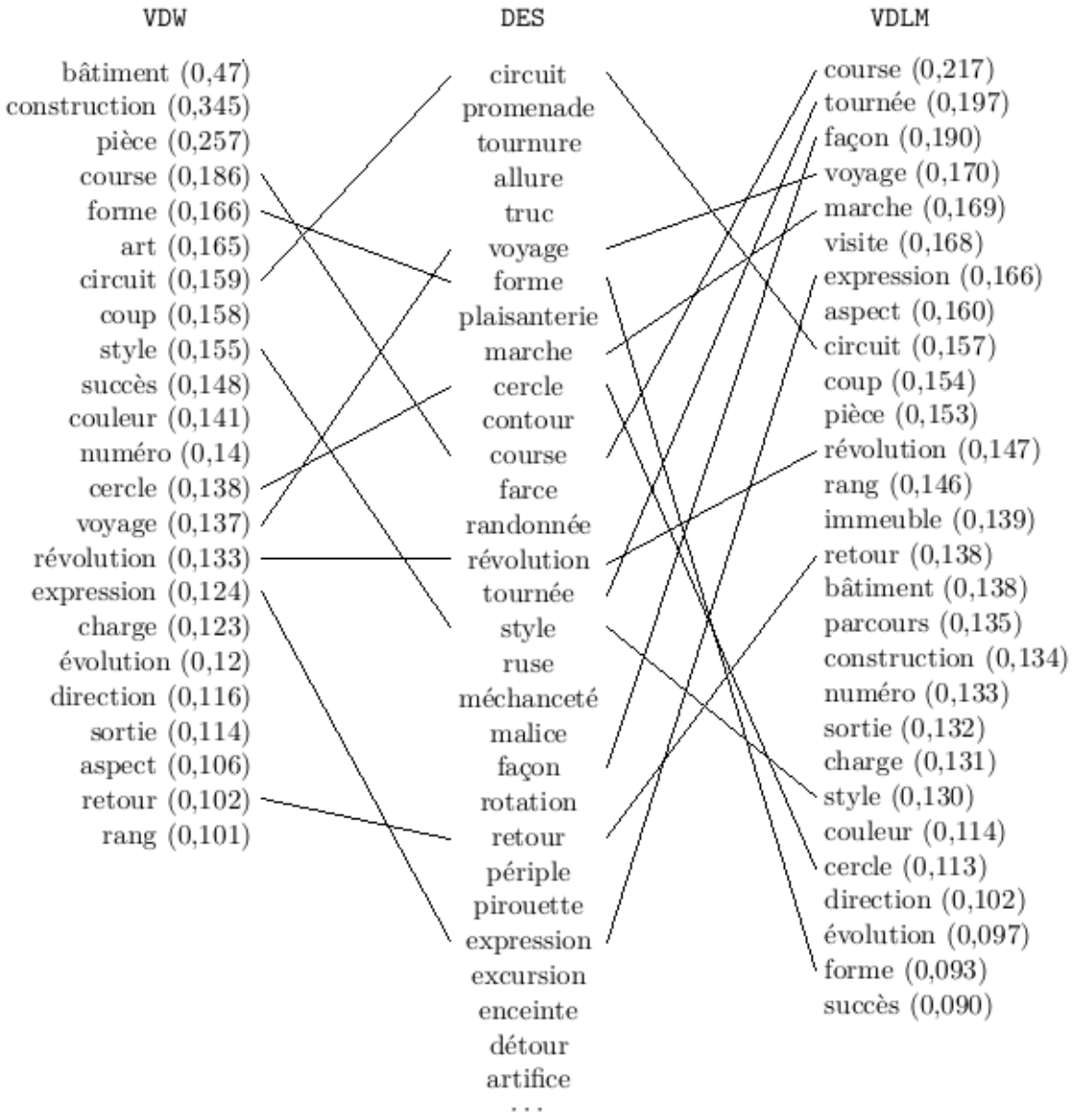


FIG. 5.3 – Illustration de la différence entre l'ordre de présentation des synonymes du nom *tour* dans le DES et le classement de ses voisins par score de Lin décroissant (VDW/VDLM).

– le classement par Lin décroissant des 23 synonymes de *tour* extraits par les VDW ;
– le classement par Lin décroissant des 29 synonymes de *tour* extraits par les VDLM.

Afin de permettre une meilleure visualisation de la différence de classement des synonymes, nous avons tracé des lignes entre les synonymes tels qu'ils sont classés par le DES (colonne centrale) et tels qu'ils sont classés par le score de Lin calculé dans les VDW et les VDLM (respectivement dans les colonnes de gauche et de droite.

Les résultats appellent plusieurs commentaires. Dans un premier temps, on remarque que parmi les 30 synonymes qui ont été classés les plus pertinents par le DES, seuls 11 ont été captés par au moins une des deux bases distributionnelles. Réciproquement, moins de la moitié des synonymes voisins figurent parmi les 30 meilleurs synonymes du DES. Ces deux constats montrent un décalage flagrant entre un classement effectué *in abstracto* et un classement qui s'appuie sur un score de similarité distributionnelle calculé à partir des usages du mot vedette en corpus.

Dans un deuxième temps, on voit que les scores de proximité calculés entre *tour* et ses synonymes dans les deux bases distributionnelle varient fortement. Afin d'observer plus nettement le décalage dans le classement des synonymes dans les deux bases de voisins, nous avons mis ces deux classements face à face dans la figure 5.4. On peut notamment observer que les synonymes qui émergent dans les VDW sont *bâtiment* et *construction*. Ils renvoient à un emploi de *tour* qui semble plus minoritaire dans le corpus Le Monde et qui est très peu privilégié par le DES, puisqu'ils apparaissent respectivement en 59^e et 90^e position dans la liste des synonymes fournie par le DES. En comparaison, les 5 meilleurs synonymes selon le score de proximité calculé dans les VDLM figurent tous parmi les 30 meilleurs synonymes du DES. De ce fait, on voit qu'ici, le filtrage opéré par les VDW met au premier plan une acception de *tour* considérée comme marginale dans le classement du DES – "construction nettement plus haute que large" –, qui favorise pêle-mêle des acceptions comme :

– "action habile et rusée commise au détriment de quelqu'un" : *plaisanterie*, *farce*, *ruse*, *méchanceté*, *malice*, etc.
– "mouvement, déplacement (à peu près) circulaire où l'on revient au point de départ" ou "périple, voyage" : *promenade*, *voyage*, *marche*, *randonnée*, *excursion*, etc.
– "manière spécifique de s'exprimer, d'être exprimé ; ce qui est exprimé" ou "manière spécifique d'être, de procéder, d'évoluer" : *allure*, *style*, *façon*, *expression*, etc.

Au delà du phénomène observé dans le cas de *bâtiment* et *construction*, les

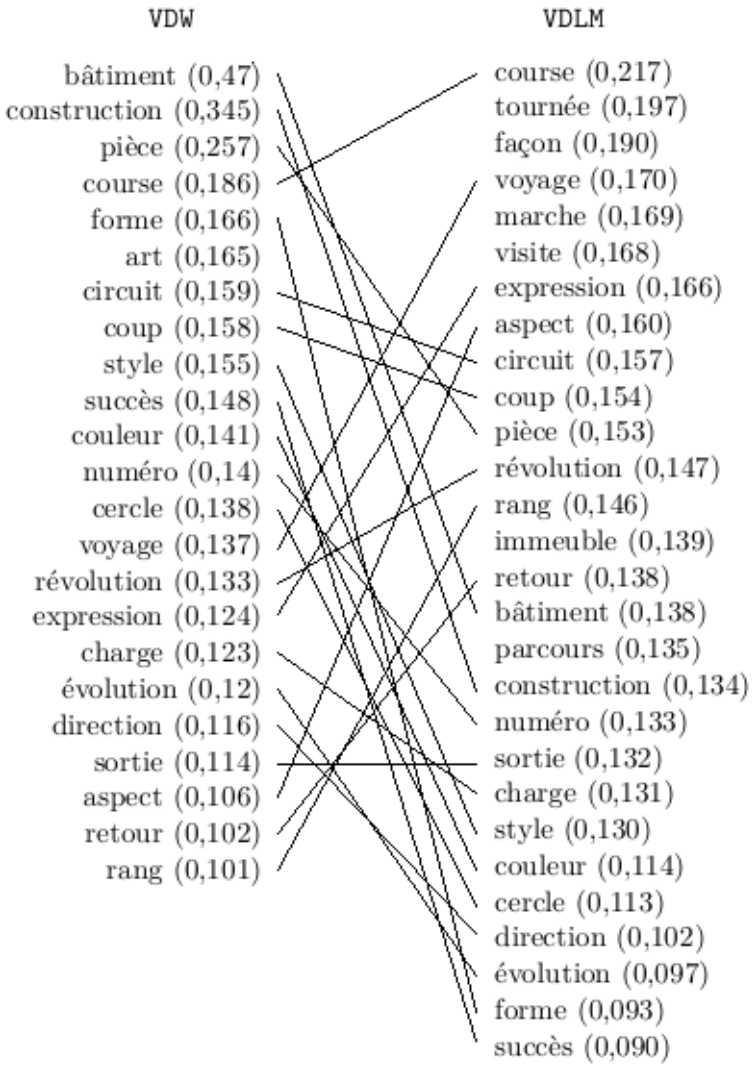


FIG. 5.4 – Illustration de la différence entre les scores de Lin calculés pour *tour* et ses synonymes dans les VDW et les VDLM.

synonymes captés par les voisins reflètent l'hétérogénéité des synonymes par le DES, qui est due à la polysémie de *tour*.

On note que certains des synonymes captés par les voisins ont été rapprochés de *tour* sur la base d'une ambiguïté. Par exemple, alors que le synonyme *pièce* renvoie à l'acception "partie d'un ensemble considérée comme un tout autonome" (l'ensemble étant ici un jeu d'échecs), l'analyse des contextes communs à *tour* et *pièce* dans le corpus Wikipédia montre que ce n'est pas cet emploi qui émerge le plus. Ces deux mots ont en effet été rapprochés sur la base de contextes comme :

- *face*_DE, *frapper*_OBJ, qui semblent renvoyer à un emploi de *pièce* comme un "petit disque plat gravé dans un métal plus ou moins précieux et servant de monnaie d'échange" ;
- *enfermer*_DANS ou *installer*_DANS, qui renvoient à l'emploi "chacune des parties d'une maison ou d'un appartement".

De la même façon, on voit que les contextes qui ont permis de rapprocher *tour* et *révolution* – *lendemain*_DE, *héros*_DE, *abandonner*_À – renvoient à l'acception de *révolution* "renversement soudain du régime politique d'une nation [...] par un mouvement populaire", qui n'est pas synonyme de *tour* (contrairement à "mouvement en courbe fermée autour d'un axe ou d'un point", à laquelle renvoie *révolution* dans la liste du DES). Ainsi, partant du fait que l'extraction de couples comme *tour/pièce* et *tour/révolution* s'appuie principalement sur des contextes d'apparition de *pièce* et *révolution* employés dans des acceptions qui ne sont pas synonymes de *tour*, on peut se poser la question de l'intérêt de ce type de rapprochements dans un système de classement des synonymes.

5.5.3 Conclusion

Après avoir recensé, dans la section 5.4, les différentes raisons pouvant expliquer la sélection – ou l'exclusion, en l'occurrence – de certains synonymes par les VDW, nous avons comparé les résultats du filtrage opéré par les VDW à celui des VDLM et des VDF. Nous avons observé une tendance globale des VDW et des VDLM à extraire les mêmes synonymes (la comparaison avec les VDF est plus difficile étant donnée la différence de taille des corpus de départ). Toutefois, pour certains mots vedettes, l'influence du corpus peut s'observer à travers le fait que les différentes bases de voisins captent des synonymes différents. Nous avons illustré ce phénomène en montrant de quelle façon les différences d'emplois des mots *doux*, *éclat* et *chasser* dans les corpus Wikipédia, Le Monde et Frantext influençaient le filtrage de leurs synonymes.

Nous avons ensuite proposé un deuxième mode d'analyse du filtrage des

synonymes qui consiste à prendre en compte leur classement par score de similarité distributionnelle décroissant. L'étude des synonymes du nom *tour* a fait entrevoir l'intérêt qu'il y a à aborder la question du filtrage du point de vue de la gradabilité : le fait qu'un même synonyme soit capté par deux ressources distributionnelles différentes ne veut pas dire que ce synonyme présente le même degré de substituabilité avec le mot vedette dans les deux corpus. Nous avons ici pu voir que le fait de prendre en compte le score de Lin dans l'observation du recouvrement entre le DES et les bases distributionnelles nous permettait d'accéder à un niveau d'analyse plus fin que l'observation de la présence/absence des synonymes dans les bases distributionnelles. Le fait de raisonner en termes de listes ordonnées complexifie toutefois la question de la similarité du recouvrement entre les bases distributionnelles : alors qu'il nous a suffi jusque-là de mesurer la présence ou l'absence d'un synonyme dans nos ressources, il s'agira ici de prendre en compte le rang auquel se situe le synonyme dans les listes comparées.

5.6 Projet d'évaluation du filtrage

Dans ce chapitre, nous avons abordé la question de ce que pourrait apporter à un dictionnaire de synonymes un filtrage par une base distributionnelle. Par l'analyse de plusieurs exemples, nous avons montré que :

1. tous les synonymes d'un mot vedette ne partagent pas ses contextes d'apparition et donc, ne sont pas extraits par les bases distributionnelles ;
2. la nature des synonymes extraits peut varier selon que l'on filtre le dictionnaire avec une ressource calculée à partir d'un corpus constitué de textes encyclopédiques, journalistiques ou littéraires.

L'idée majeure qui a parcouru ce chapitre est que la pertinence des synonymes d'un mot donné n'est pas absolue mais varie en fonction du type de corpus dans lequel le mot est employé. Nous avons donc proposé d'exploiter les informations que fournit l'ADA sur la distribution du mot dont on recherche le synonyme. Il s'agirait alors de se servir de ressources distributionnelles en support du DES pour savoir quels sont les synonymes qui sont les plus pertinents pour un type de texte donné et de réorganiser les résultats fournis en conséquence.

Cette question de la pertinence reste ici en suspens. En effet, à l'heure actuelle, nous n'avons pas encore procédé à l'évaluation des apports du filtrage. Nous avons donc prévu de poursuivre la présente étude en mettant en place un protocole dans lequel il sera demandé à des utilisateurs de sélectionner

les meilleurs synonymes pour un mot donné dans un ensemble de phrases extraites des corpus Wikipédia, Le Monde et Frantext. Nous faisons ainsi l'hypothèse que le choix des utilisateurs se portera en priorité sur les synonymes qui sont voisins du mot cible dans la base distributionnelle calculée à partir du corpus d'où a été extraite la phrase.

Les modalités du protocole sont encore à définir, aussi bien au niveau de la formulation des consignes que du choix de l'interface de sélection des synonymes. Nous avons pour le moment envisagé trois types de formulaires à présenter aux participants. Le premier, présenté à la figure 5.5, consiste à classer les n synonymes les plus pertinents pour un mot figurant en gras dans une phrase donnée. Le deuxième – 5.6 – implique un choix binaire oui/non (avec une possibilité de réponse neutre) sur l'ensemble des synonymes. Dans le troisième – 5.7 –, il est demandé au participant de choisir l'énoncé qui correspond le mieux à son intuition sur le degré de substituabilité entre le mot en gras et ses synonymes. Chacun de ces protocoles entraîne des conséquences sur la nature des résultats fournis. Dans la mesure où nous n'avons pas encore mesuré l'étendue de ces conséquences, nous remettons à une étude ultérieure la question de savoir lequel de ces modes de présentation serait le plus adapté au phénomène que nous cherchons à observer.

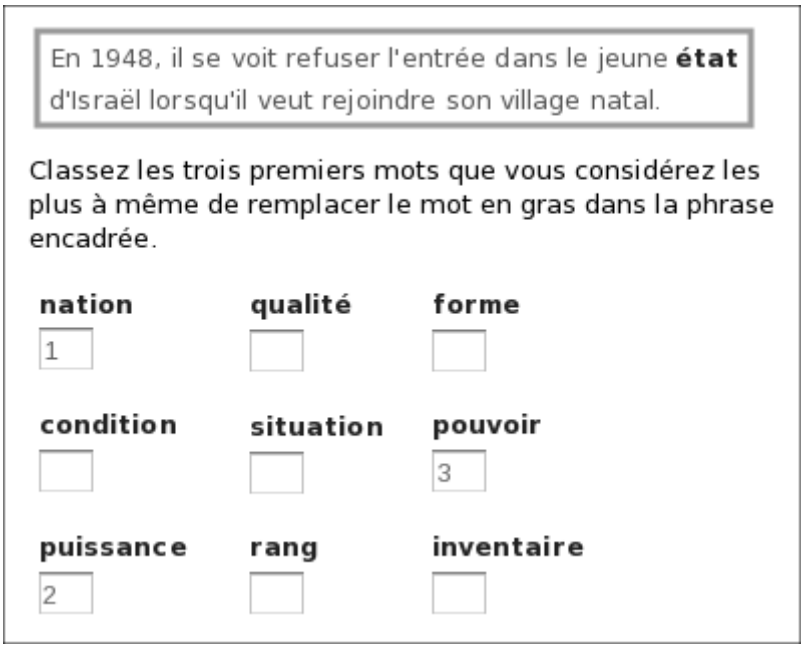


FIG. 5.5 – Proposition d'interface n° 1.

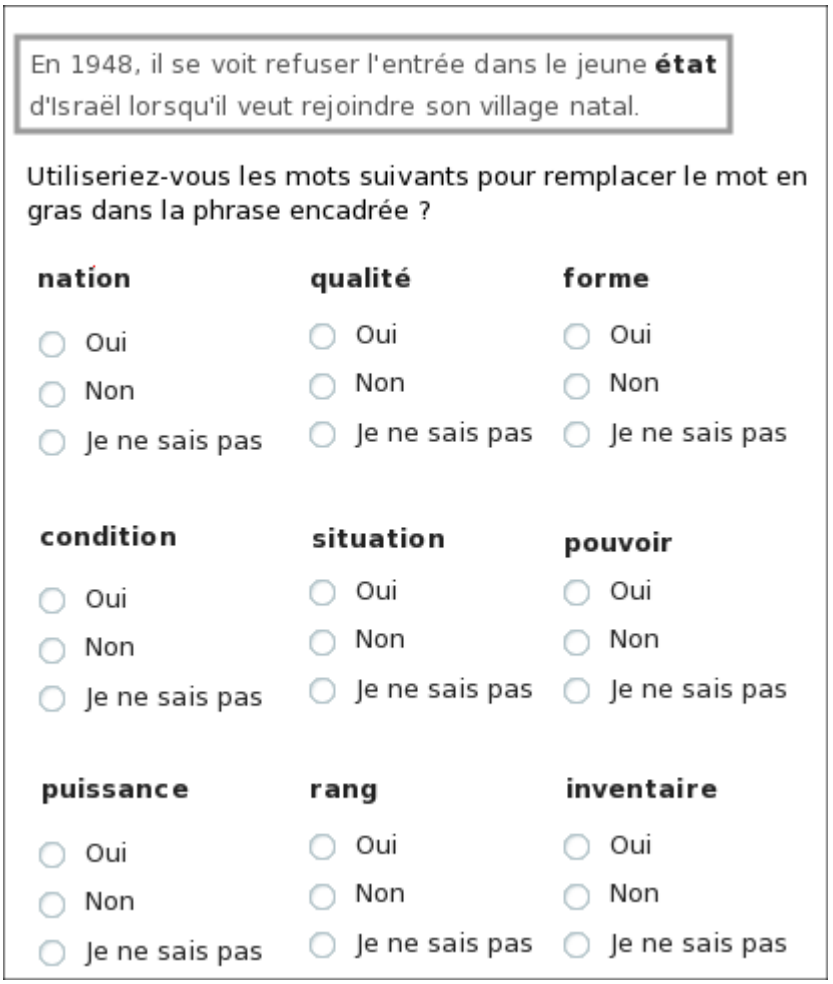


FIG. 5.6 – Proposition d'interface n° 2.

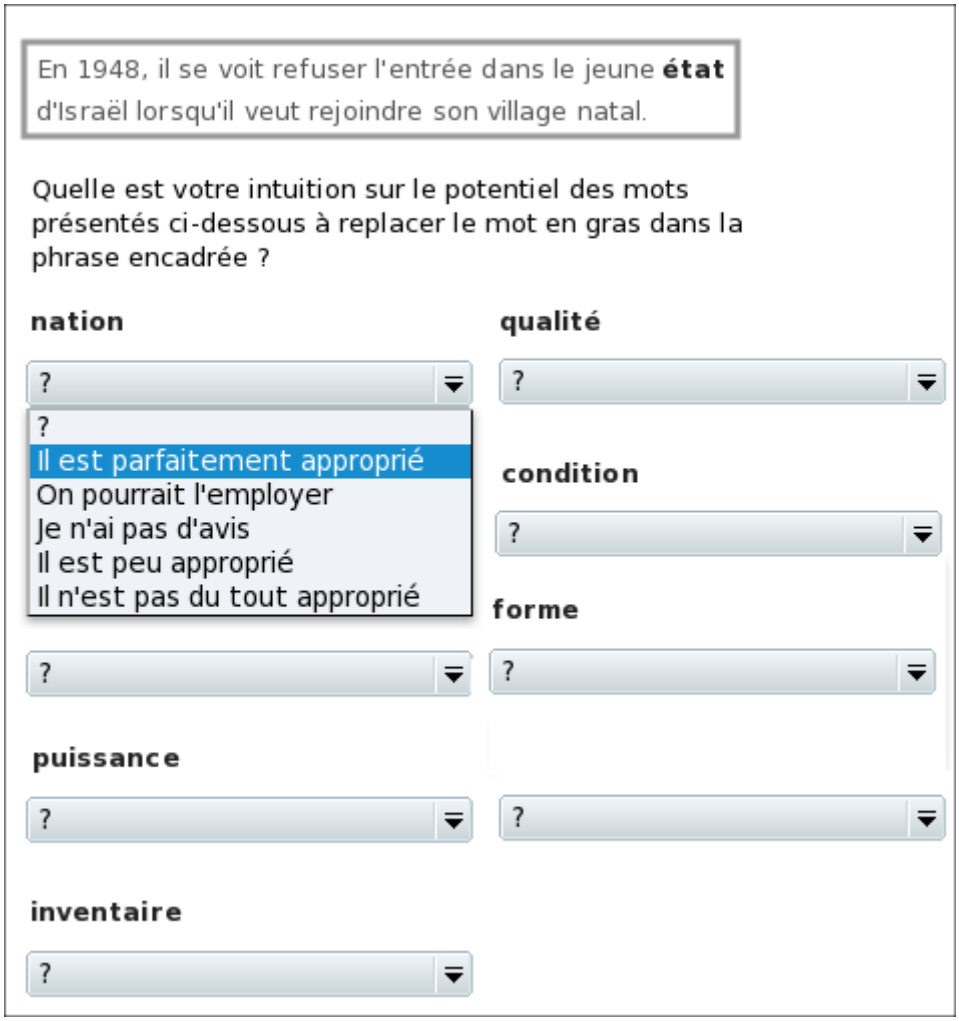


FIG. 5.7 – Proposition d'interface n° 3.

Chapitre 6

Une description à la fois syntagmatique et paradigmatique de l'antonymie

Sommaire

Dans ce chapitre, nous abordons la question de l'apport que peut représenter une base distributionnelle pour la description linguistique et l'extraction de la relation d'antonymie. Une série de travaux récents à l'interface entre sémantique lexicale, discours et psycholinguistique tendent à montrer que la relation antonymique possède la propriété de se manifester à la fois sur les plans paradigmatique et syntagmatique. Paradigmatique, bien sûr, parce que

deux antonymes s'opposent à l'intérieur d'un champ sémantique déterminé et se conforment donc au principe de substituabilité. Syntagmatique aussi parce que deux mots considérés comme antonymes tendent à apparaître ensemble dans des relations aussi bien inter qu'intrapropositionnelles et plus particulièrement dans des constructions contrastives, comme dans les exemples suivants extraits du corpus Wikipédia :

(1) Suivant les scènes, il est *à la fois* le **bourreau** *et* la **victime**, le poursuivant et le poursuivi, etc.

(2) Kant défend donc un agnosticisme métaphysique : on peut [sic] en réalité *ni* **rejeter** *ni* **approuver** ses affirmations.

(3) Pour les chrétiens, **catholiques** *ou* **protestants**, Jésus de Nazareth s'est sacrifié lui-même pour sauver le genre humain.

Cette capacité des couples antonymiques à fonctionner simultanément sur ces deux plans est remarquable car elle n'est pas partagée par les autres relations lexicales (synonymie, hyperonymie, co-hyponymie), pour lesquelles le plan paradigmatique prévaut.

Partant de ce constat d'un double fonctionnement, on peut faire l'hypothèse que le fait, pour un couple de mots sémantiquement opposés, de fonctionner à la fois sur le plan paradigmatique et syntagmatique est un indice d'un degré élevé d'antonymie. Nous avançons également le fait que cette propriété peut être exploitée pour l'extraction des couples d'antonymes, dont l'intérêt se fait sentir dans un domaine comme l'analyse automatique d'opinions ou pour l'identification de segments discursifs entretenant une relation de contraste (Marcu et Echihabi, 2002). D'une part, on sait le potentiel des méthodes distributionnelles à capter des couples d'antonymes. Leur repérage dans la masse des résultats fournis reste cependant problématique. D'autre part, des travaux comme ceux de Jones (2002) et Lobanova *et al.* (2010) mettent en œuvre des méthodes inspirées de Hearst (1992) qui consistent à construire ou extraire des patrons lexico-syntaxiques et à les projeter sur un corpus afin de faire émerger les couples entretenant une relation donnée. Toutefois, alors que Hearst, cherche à extraire les couples entretenant une relation d'hyperonymie – avec des patrons de type *X including Y* –, Jones, Lobanova et collègues visent l'extraction d'antonymes en s'appuyant sur des constructions comme *à la fois X et Y*, *ni X ni Y* ou encore *X ou Y*, dont nous avons rapporté quelques exemples dans le paragraphe ci-dessus.

L'originalité de notre approche réside donc dans le fait qu'elle prend en considération les propriétés à la fois syntagmatiques et paradigmatiques des couples d'antonymes. Elle consiste en effet à croiser :

– les résultats fournis par la projection sur le corpus Wikipédia d'une série

de patrons configurés pour repérer des constructions antonymiques ;
– les voisins de Wikipédia (VDW), une base distributionnelle calculée à partir de ce même corpus.

Ce principe d'un filtrage mutuel des couples extraits par les patrons et par les voisins – et inversement – nous permet ainsi d'accéder aux couples qui sont à la fois cooccurrents et substituables, dont on suppose qu'il auront le plus de chances d'être considérés comme de *bons* antonymes par des locuteurs (nous verrons plus loin quels sont les critères qui définissent ces *bons* antonymes, que nous appellerons *canoniques*).

À la section 6.1, nous faisons le point sur les différents travaux qui ont mené les lexicologues à passer d'une description purement sémantique de l'antonymie à une approche psycholinguistique épaulée par des méthodes de linguistique de corpus. Nous développons, à la section 6.2, la démarche que nous avons évoquée ci-dessus, puis, à la section 6.3, nous décrivons les méthodes que nous avons choisies pour l'évaluer, avant de présenter les résultats que nous avons obtenus.

6.1 Décrire la relation d'antonymie

Parmi les travaux qui se sont penchés sur l'antonymie, on peut distinguer plusieurs types d'approches. Une première consiste à aborder la relation du point de vue de ses propriétés sémantico-logiques (Lyons, 1977; Cruse, 1986). Elle a mené à la création de typologies de l'antonymie comme celle de Cruse (2004), que nous commentons dans la section 6.1.1. Le fait de définir l'antonoymie sur des critères purement sémantiques montre vite ses limites : les classifications ne permettent pas d'expliquer le fait que le caractère antonymique se ressente davantage pour certains couples de mots opposés que pour d'autres (section 6.1.2). De plus, elles sont établies *in abstracto* et négligent donc l'emploi qui est fait des antonymes en contexte. Nous évoquons ainsi à la section 6.1.3 une série de travaux qui montrent que les antonymes présentent une tendance à apparaître dans les mêmes phrases, et plus particulièrement dans des constructions qui jouent un rôle au niveau discursif.

6.1.1 Typologies sémantico-logiques

De toutes les relations lexicales, l'antonymie est certainement celle qui a fait l'objet du plus grand nombre de classifications, plus ou moins complexes et souvent différentes d'un auteur à l'autre. En cela, l'antonymie se distingue de la synonymie qui, comme nous l'avons vu dans le chapitre précédent, ne se divise qu'en deux catégories relativement consensuelles, à savoir la synonymie

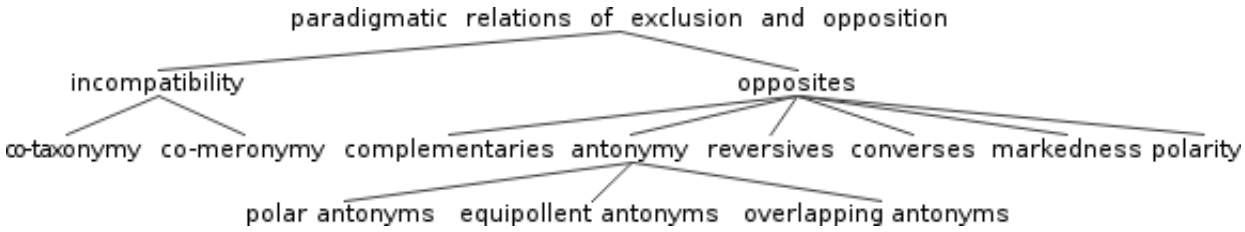


FIG. 6.1 – Classification des relations d'opposition et d'exclusion de Cruse (2004).

parfaite/absolue et la quasi-synonymie/parasynonymie. Ainsi, Cruse (2004) définit pas moins de 10 relations d'opposition ou d'exclusion. Il les distingue en deux groupes de haut niveau, à savoir les relations d'incompatibilité et les relations d'opposition.

Les relations d'incompatibilité – ou *incompatibilités multiples*, dans la terminologie de Lyons – portent sur des ensembles non binaires comme les quatre saisons, les jours de la semaine ou les grades de l'armée. Autrement dit, des co-hyponymes. Cruse classe aussi dans cette catégories les co-méronymes comme *œil*, *bouche*, *nez* ou *oreille*, qui sont des méronymes de *visage*.

Les relations d'opposition sont au nombre de 6. Elles renvoient aux différents rapports sémantico-logiques que peuvent entretenir deux mots de sens opposés dans un ensemble binaire. On peut voir que parmi ces relations figure l'antonymie, qui renvoie donc ici à un sous-type de la relation d'opposition et non comme la relation d'opposition elle-même. En effet, on distingue traditionnellement deux acceptions du terme *antonymie*. La notion d'antonymie au sens large peut être définie de la façon suivante :

> In its broadest sense, antonymy covers a wide range of word pairs expressed by different part-of-speech categories as long as these words express the opposite of each other. (Lobanova, 2012, p. 20)

Au sens restreint, celui de Cruse (2004), elle désigne une sous-classe qui englobe trois relations, la principale étant l'antonymie polaire. Cette relation, également appelée *contraire* ou *gradable*, est considérée par des auteurs comme Cruse (2004) ou Lyons (1977) comme la *véritable* antonymie. Voici quelques-unes des propriétés de deux antonymes polaires comme *chaud* et *froid* :

- ils constituent les pôles d'une échelle de valeurs donnée (ici, la température) ;
- ils sont gradables et l'on peut ainsi désigner des points intermédiaires sur l'échelle en les situant par rapport à ces pôles soit :

  - en utilisant des adverbes intensificateurs (*assez chaud*, *un peu froid*) ;
  - en niant les deux prédicats contraires (*Ce n'est ni chaud ni froid*). Ces valeurs intermédiaires peuvent être lexicalisées (*tiède/frais*) ;
- les pôles ne constituent que des valeurs de l'échelle lexicalisées, et ne définissent en rien des frontières (*très chaud/très froid*, *brûlant/glacial*).

Cette description étant trop spécifique pour l'usage que nous comptons faire du concept d'antonymie, nous employons par la suite ce mot au sens large.

D'autres typologies, comme celles que l'on peut trouver dans des manuels de lexicologie (Lehmann et Martin-Berthet, 2011), ont tendance à fusionner certaines des catégories dégagées par Cruse. Ainsi, elles opposent traditionnellement les antonymes polaires aux *complémentaires* (ou *contradictoires*) comme *vivant/mort*. Ils se distinguent principalement des polaires du fait qu'ils ne sont pas gradables. Les antonymes contradictoires divisent un domaine en deux parties mutuellement exclusives : l'affirmation de l'un implique la négation de l'autre, et inversement (principe du tiers exclu).

Parmi les relations d'opposition récurrentes dans la littérature, on peut également évoquer l'antonymie *converse* ou *réciproque*. Cette relation porte sur deux mots qui expriment les deux facettes d'une même action (*prêter/emprunter*), d'un lien spatio-temporel (*avant/après*) ou d'une relation sociale (*maître/serviteur*) ou familiale (*oncle/neveu*). Ils ont en commun le fait que la substitution de l'un par l'autre implique une inversion des actants :

(4) Pierre a **prêté** un bic à Paul. ⇒ Paul a **emprunté** un bic à Pierre.

Toutefois, ces typologies présentent des limites. Des auteurs comme Murphy (2003) critiquent notamment le fait qu'elles soient établies *in abstracto* et qu'elles se trouvent donc en décalage avec les *véritables* emplois des antonymes. Ces auteurs privilégient alors une étude de la relation basée sur une démarche psycholinguistique. Nous décrivons ces approches dans la section suivante.

6.1.2 Limites du critère sémantique

La principale critique adressée aux classifications sémantico-logiques est qu'elles sont déconnectées des emplois effectifs. Par exemple, Lobanova (2012) montre qu'il est possible de rencontrer des emplois gradables d'antonymes contradictoires. Ainsi, elle montre qu'il est possible de trouver des occurrences de *très marié* ou *très célibataire* alors que le couple *marié/célibataire* est considéré comme non gradable (cette critique peut toutefois être nuancée par le fait que la gradation produit un glissement de sens). De la même façon, elle relativise la binarité des contradictoires en prenant l'exemple des

zombies, qui ne sont ni vivants ni morts, et des escargots, qui ne sont ni mâle ni femelle. Murphy (2003) critique également la typologie de Cruse sur plusieurs points :

- on peut aisément trouver des couples d'antonymes qui n'appartiennent à aucune classe comme *demander/répondre* (ces mots ne renvoient pas à deux facettes d'une même action mais à deux actions distinctes) ;
- la typologie ne permet pas d'expliquer pourquoi, parmi les quatre saisons, le couple *été/hiver* s'oppose davantage que *automne/printemps*.

Ainsi, les typologies et les critères de la sémantique en général trouvent leurs limites lorsqu'il s'agit d'expliquer les différences de jugement entre les paires de mots de sens opposés.

Les approches cognitives se sont alors focalisées sur l'aspect perceptif, en montrant que le jugement d'antonymie n'était pas binaire mais graduel. Elles ont notamment développé le concept de *canonicité* qui distingue parmi les paires d'antonymes celles qui sont immédiatement reconnues comme antonymiques – *gentil/méchant* – alors que le jugement est plus mitigé pour d'autres – *gentil/mauvais*. Cette conception de l'antonymie s'appuie notamment sur l'idée selon laquelle, contrairement à une relation comme la synonymie, l'antonymie a un fonctionnement de type *one-to-one*. Autrement dit, il existe certains facteurs qui font que, lorsqu'on présente une série de mots à des locuteurs dans une tâche d'association, ces derniers ont tendance à produire les mêmes antonymes (alors que ce phénomène est plus variable pour d'autres relations (Deese, 1964)). Jones *et al.* (2012, p. 1) montrent que cette perception de l'antonymie se reflète à travers l'usage qui est fait des termes *opposite* et *synonym* dans le Corpus of Contemporary American English (Davies, 2010) : le premier s'emploie majoritairement précédé d'un article défini alors que le deuxième s'emploie précédé d'un indéfini. Nous avons adapté cette expérience sur le français en comparant les fréquences des résultats fournis par Google pour les requêtes `le/l' * de` et `un * de`, l'astérisque étant successivement remplacé par les noms *opposé*, *contraire* et *antonyme*, puis par *synonyme*. Les résultats, rapportés au tableau 6.1, vont dans le sens des observations de Jones *et al.* (2012) et donc dans celui d'une conception *one-to-one* de l'antonymie.

La question de la canonicité a été également étudiée par Murphy (2003, 2006), qui met en exergue le rôle de facteurs non sémantiques dans le jugement du degré d'opposition de deux mots. Le fait que l'un des antonymes soit un dérivé morphologique de l'autre est l'un de ces facteurs : l'oppostion portée par le couple *edible/inedible* est ainsi perçue comme davantage canonique que le couple *edible/uneatable*. Plus généralement, il semble que toute similarité dans la forme des mots opposés ait une influence favorable sur leur propension à être reconnus comme des antonymes canoniques (*awake/asleep*

		opposé	*contraire*	*antonyme*	*synonyme*
`fréquence brute`	*le*	11,000,000	25,000,000	152,000	779,000
	un	34,400	117,000	22,000	2,370,000
`%`	*le*	99,7 %	99,5 %	87,4 %	24,7 %
	un	0,3 %	0,5 %	12,6 %	75,3 %

TAB. 6.1 – Comparaison de la fréquence des emplois des déterminants *le* et *un* avec les noms *opposé*, *contraire*, *antonyme* et *synonyme* dans Google.

vs *up*/*asleep*). Cette influence de facteurs non sémantiques poussera Murphy à envisager l'antonymie comme une relation davantage lexicale que sémantique.

6.1.3 Approche linguistique de corpus

La question de savoir ce qui fait que deux mots sont immédiatement reconnus comme des antonymes alors que le jugement est plus mitigé pour d'autres a poussé les psycholinguistes à s'intéresser aux occurrences des antonymes dans les textes et donc à adopter des approches directement issues de la linguistique de corpus :

> A combination of corpus data, elicitation data and judgement data is valuable in order to determine if and how antonym word pairs vary in canonicity. It also sheds light on different aspects of the issue. [...] we believe that a methodologically sound descriptive study of linguistics is cyclic and preferably includes both corpus evidence and intuitive data (psycho-linguistic experimental data). (Willners et Paradis, 2010, p. 5)

Le phénomène de canonicité a été étudié pour la première fois par le psychologue James Deese (1964, 1965). Il observe que, dans des tests d'association, les locuteurs ont une forte tendance à produire des antonymes lorsqu'un adjectif leur est présenté. Il émet alors l'hypothèse selon laquelle c'est la relation d'antonymie qui structure les adjectifs dans le lexique mental.

Il remarque également que seul un sous-ensemble de ces couples présente la propriété suivante : pour un couple A/B, B est la réponse la plus fréquemment fournie lorsque le stimulus est A, et inversement. Cette réciprocité ne caractérise pas l'ensemble des couples, ce qui pousse Deese à postuler une différence dans le degré d'opposition de deux adjectifs : la réciprocité des couples traduit alors selon lui un degré d'opposition élevé, qu'il explique par le fait que ces mots partagent les mêmes contextes. Autrement dit, le degré

d'antonymie ressenti entre deux adjectifs varierait en fonction des noms qu'ils modifient :

> Even the apparently synonymous pairs, *big/little* and *large/small*, have different uses in the language [...]. Think of the difficulty, for example, of substituting *large* for *big* in "He opened his big mouth". Therefore, it is not surprising that much of the meaning of one of these pairs, as the meaning is described by linguistic context, cannot be determined from the other. (Deese, 1965, p. 127)

Deese fait ainsi une corrélation entre la perception des couples d'antonymes et les emplois qui en sont faits dans les textes.

Cette intuition préfigure l'hypothèse de la substituabilité formulée par Charles et Miller (1989). En effet, dans le cadre du projet WordNet, ces derniers ont poursuivi les travaux de Deese en cherchant à démontrer une relation entre le degré d'antonymie de deux mots et leur propension à apparaître dans les mêmes contextes. Pour cela, ils extraient du Brown Corpus (Francis et Kucera, 1979) 25 phrases contenant l'adjectif *strong* et 25 autres contenant son antonyme *weak*. Ils effacent ensuite l'adjectif de chaque phrase et demandent à des locuteurs de remplir les blancs avec l'un des adjectifs. Les résultats montrent que les locuteurs ne confondent les contextes qu'à de rares occasions (ce qui est assez peu étonnant étant donné que c'est l'ensemble de la phrase qui conditionne le choix de l'adjectif). La même expérience a ensuite été menée avec les antonymes *public/private* et des contextes réduits à des syntagmes nominaux (l'hypothèse étant qu'un contexte plus restreint aurait moins de chance d'orienter les sujets dans le choix des adjectifs manquants). Dans les deux cas, le taux de confusion entre les contextes des paires antonymiques reste très faible. Ces résultats poussent Charles et Miller à conclure que tous les noms modifiables par un adjectif donné ne le sont pas forcément par son antonyme (pour reprendre l'exemple de Lobanova (2012), il y a plus de chances que le nom *café* soit modifié par *fort* que par *faible*). D'après leurs expériences, la théorie de la substituabilité ne se vérifie donc pas. Il faut toutefois garder à l'esprit que la taille réduite de l'échantillon observé rend la généralisation des phénomènes observés assez délicate : des travaux comme ceux de Grefenstette (1992a) et Mohammad *et al.* (2013) réfutent ces résultats.

Charles et Miller se tournent alors vers l'hypothèse d'une corrélation entre le degré d'antonymie perçu entre deux mots et leur tendance à cooccurrer dans les mêmes phrases. Ils comparent ainsi dans le Brown Corpus le nombre de cooccurrences de *big* et *little*, puis de *large* et *small*. Ces deux couples présentent une relation antonymique manifeste, qu'ils qualifient de *directe* (ou *canoniques*, cf. section 6.1.2). Ils comptent ensuite le nombre de cooccur-

	paire	occurrence effective	occurrence théorique	rapport
antonymes directs	big/little	12	1,6	7,5
	large/small	26	3,2	8
antonymes indirects	large/little	3	1,7	1,8
	big/small	4	2,9	1,4

TAB. 6.2 – Résultats des mesures de cooccurrences de Charles et Miller (1989).

rences de *large* et *little* puis de *big* et *small* dans le Brown Corpus. Les couples *large/little* et *big/small*, en revanche, sont composés de mots ayant des sens opposés mais, du moins pour un locuteur natif, ils n'apparaissent pas intuitivement comme des couples présentant une relation antonymique canonique. Charles et Miller les qualifient d'*indirects*. Les résultats, rapportés dans le tableau 6.2 montrent clairement que les termes considérés *a priori* comme antonymes directs présentent une tendance plus élevée à la cooccurrence que les antonymes indirects. Ces données tendent à vérifier l'hypothèse de la cooccurrence. Les travaux de Charles et Miller (1989) posent ainsi les bases d'une approche psycholinguistique de l'antonymie soutenue par des données issues des corpus.

L'hypothèse de la cooccurrence a ensuite été vérifiée par Justeson et Katz (1991), qui mettent en lumière les structures dans lesquelles cooccurrent les antonymes. Fellbaum (1995) montrera que cette tendance à la cooccurrence se vérifie également pour les couples de mots opposés appartenant à des catégories grammaticales différentes comme la forme verbale *increase* et le nom *decrement*.

Les travaux de Jones (2002) et Jones *et al.* (2012), que nous avons évoqués à la section 4.1.2, s'inscrivent directement dans la lignée de Fellbaum (1995) : non seulement ils confirment la tendance qu'ont les antonymes à cooccurrer, mais ils montrent également que ces cooccurrences se font dans des constructions particulières qui portent des valeurs discursives. À travers l'étude des contextes d'apparition d'une série d'antonymes, ils mettent ainsi au jour 8 types de configurations. On peut par exemple citer :

– l'antonymie coordonnée, qui consiste à exprimer une notion d'exhaustivité en mettant en relation les antonymes au moyen d'une conjonction de coordination :

(5) But assuming no scandals, **old** *or* **new**, precipitate presidential disgrace, what is he to do if a triumphal place in history is to

be assured ?

– l'antonymie transitionnelle, qui traduit le changement d'un état à un autre :

(6) Her film career similarly has lurched *from* **success** *to* **failure**, with enormous period out of work.

– l'antonymie négative, qui consiste à nier l'un des antonymes de la paire afin de renforcer le sens de l'autre :

(7) Well, without the combination of an arms race and a network of treaties designed for **war**, *not* **peace**, it would not have started.

Cette série d'études montre ainsi le cheminement qui a été effectué entre une approche psycholinguistique de l'antonymie et une démarche de linguistique de corpus. Il en ressort une conception de l'antonymie comme une relation qui se manifeste – et se construit – sur le plan syntagmatique, ce qui rompt avec une vision traditionnelle d'une relation qui fonctionne uniquement sur le mode de la substituabilité (le caractère novateur de ce point de vue est notamment illustré par le titre de l'étude de Murphy (2006) – *Antonyms as lexical constructions : or, why paradigmatic construction is not an oxymoron*). C'est donc sur ce constat d'un fonctionnement syntagmatique de l'antonymie que nous appuyons notre hypothèse selon laquelle le fait de croiser une base distributionnelle et des patrons lexico-syntaxiques – basés sur ceux de Jones (2002) – nous permettra d'extraire des antonymes canoniques.

6.2 Combiner deux modes de repérage de l'antonymie

Nous abordons ici la question de la canonicité en nous appuyant sur les propriétés des antonymes que sont la cooccurrence et la substituabilité. Le fait de disposer d'un corpus, de patrons et d'une ressource distributionnelle nous permet en effet d'observer à grande échelle le comportement des couples antonymiques du point de vue paradigmatique et syntagmatique. Nous aurions pu faire le choix de poursuivre l'approche des travaux de Murphy, Jones et leurs collègues, en prenant comme eux en entrée de nos traitements une liste déjà constituée de couples antonymiques – extraite de JeuxDeMots, par exemple – afin d'observer leur tendance à opérer sur ces deux plans. Nous avons toutefois privilégié une approche inductive qui a consisté à utiliser des couples qui ont été extraits – soit par les patrons, soit par l'analyse distri-

butionnelle automatique (ADA) – à partir de la même source de données, à savoir le corpus Wikipédia.

À notre connaissance, cette démarche n'a été adoptée que dans Lin *et al.* (2003). Leur étude part du constat qu'une base distributionnelle construite en suivant la méthode de Lin (1998a), qui s'appuie sur les contextes syntaxiques des mots (comme les voisins), et qui est donc censée capter en priorité les relations de similarité (cf. section 4.1), extrait également des couples d'antonymes et de co-hyponymes. Le but est alors d'écarter ces couples en les filtrant à l'aide de deux patrons censés capter les relations d'opposition, à savoir *from X to Y* et *either X or Y*. Afin de montrer que cette méthode permet de discriminer efficacement les antonymes des synonymes, ils mettent en place un protocole qui consiste à :

1. sélectionner un ensemble de 160 couples de synonymes et d'antonymes extraits d'un thésaurus ;
2. calculer, pour chacun d'eux, le rapport entre le nombre de résultats renvoyés par un moteur de recherche pour :
   – la requête *X NEAR Y*, soit l'ensemble des cas où les deux mots apparaissent à proximité l'un de l'autre ;
   – les deux mots du couple placés en position X et Y – puis Y et X – dans les patrons.

   Les couples pour lesquels le rapport ne dépasse pas un certain seuil sont alors considérés comme entretenant une relation d'opposition ou d'incompatibilité[1].

Les résultats fournis sont plus que satisfaisants puisque cette méthode obtient un rappel de 86,4 % et une précision de 95 %.

Toutefois, si cette méthodologie est proche de celle que nous mettons en place ici, le but est différent puisque notre étude vise en premier lieu l'observation des propriétés linguistiques de l'antonymie. Nous décrivons dans cette section les méthodes que nous employons pour aborder les aspects paradigmatique – section 6.2.1 – et syntagmatique – section 6.2.2 – de la relation.

6.2.1 Plan paradigmatique : analyse distributionnelle

Nous avons vu dans le chapitre 4 que l'antonymie faisait partie des relations qu'il est possible d'identifier parmi les paires de voisins. Deux mots qui entretiennent une relation d'antonymie partagent en effet beaucoup de leurs contextes d'apparition. Ainsi, dans les VDW, *entrée* et *sortie* apparaissent dans des contextes comme *sas*_DE, *périphérique*_DE ou *impédance*_DE. Cela

[1] Ils mettent également en œuvre une méthode pour capter les synonymes basée sur un dictionnaire bilingue que nous ne décrivons pas ici.

est dû au fait que, paradoxalement, les antonymes – en particulier lorsqu'ils sont canoniques – partagent la plus grande partie de leur sens (Cruse, 1986; Murphy, 2003). Par exemple, les mots *géant* et *nain* partagent un nombre important de propriétés – ils renvoient à des êtres vivants, humains, qui ont deux bras, deux jambes, qui peuvent parler, etc. – et ne se distinguent que par la valeur de la variable – ou *dimension* – TAILLE. Réciproquement, il est difficile de trouver une dimension sur laquelle opposer deux mots qui ne partagent aucune propriété comme *géant* et *épistémologie*. Ce fonctionnement paradigmatique de l'antonymie a été notamment décrit par Lyons (1968) ou Cruse (1986).

La première étude qui a mis en lumière la présence d'antonymes dans les bases distributionnelles est celle de Grefenstette (1992a). En croisant une ressource générée par son système Sextant et un sous-ensemble du jeu d'antonymes construit par Deese, il démontre la tendance des mots de sens opposés à apparaître dans des contextes identiques. Dans une étude plus récente, Mohammad *et al.* (2013) renouvellent plusieurs expériences parmi celles que nous avons évoquées plus haut à l'aide de méthodes et ressources alors inaccessibles à l'époque comme le Turc mécanique ou le corpus N-gram de Google. En s'appuyant sur le British National Corpus et un jeu de 1358 couples de mots "fortement contrastifs" extraits de WordNet, ils vérifient l'hypothèse de la substituabilité : le score de Lin (1998a) moyen calculé pour les antonymes est significativement plus élevé que pour un jeu de couples constitués de façon aléatoire. Plus étonnant, ils montrent que cette tendance est significativement plus marquée que pour un jeu de synonymes également extrait de WordNet.

Dans notre expérience, nous utilisons les VDW comme base distributionnelle.

6.2.2 Plan syntagmatique : patrons lexico-syntaxiques

Les approches à base de patrons sont bien connues du TAL. Initiées par Hearst (1992) dans le cadre du repérage de la relation d'hyperonymie, elles s'appuient sur le fait que deux mots qui entretiennent une relation sémantique donnée peuvent apparaître dans des contextes où le lien entre les deux mots est explicite. En cela, elles se distinguent des approches distributionnelles, qui infèrent des relations de sens sans que ces dernières ne soient forcément exprimées. Les patrons utilisés peuvent être soit définis au préalable par l'observation des données ou par introspection, soit extraits du corpus par la projection d'un jeu de couples portant la relation souhaitée qui auront été extraits d'une ressource de référence. On parle dans ce dernier cas d'*amorçage*, ou *bootstrapping* (Nazarenko, 2004).

patron	exemple
X *ou* **Y**	Cette connexion peut être **temporaire** *ou* **définitive**.
X *comme* **Y**	Elle peut **s'arrêter** *comme* **continuer**.
à la fois **X** *et* **Y**	Sa production est *à la fois* **fermière** *et* **industrielle**.
de/depuis **X** *à/jusqu'à* **Y**	La région s'étend *du* **nord** *au* **sud** sur 350 km
entre **X** *et* **Y**	[Il] ne voit aucun conflit *entre* **science** *et* **religion**
plus/plutôt/moins/ autant/aussi (bien) **X** *que* **Y**	Il se déguste *aussi bien* **chaud** *que* **froid**
X *plutôt que* **Y**	Pourquoi est-il une **musique** *plutôt que* du **bruit** ?
soit **X** *soit* **Y**	Les coups francs sont *soit* **directs** *soit* **indirects**.
ni **X** *ni* **Y**	Il n'est donc *ni* **explicite** *ni* **implicite**

TAB. 6.3 – Patrons retenus pour la projection sur corpus.

Dans cette étude, nous avons choisi de tirer parti des descriptions réalisées sur l'antonymie et de nous appuyer sur les constructions antonymiques que Jones (2002) met au jour à partir des contextes d'apparition d'une liste de paires d'antonymes, selon une méthode initiée par Fellbaum (1995). Notre approche consiste donc à projeter les patrons lexico-syntaxiques identifiés par Jones comme caractéristiques de l'antonymie sur l'ensemble du corpus Wikipédia. Nous cherchons ensuite à évaluer le caractère antonymique de l'intersection des couples extraits par ces patrons et de ceux qui ont été extraits par l'ADA. Pour cela, nous nous référons dans un premier temps à un dictionnaire d'antonymes, puis nous mettons en place une expérience visant à recueillir l'intuition de locuteurs sur ces paires.

Les études de Jones (2002) ayant été menées pour l'anglais, nous avons adapté ses patrons au français, obtenant ainsi la liste donnée dans le tableau 6.3. Traduites en patrons, certaines de ces constructions sont ambigües (en particulier *X ou Y*). Étant donnée la taille du corpus considéré – le corpus Wikipédia –, notre objectif a été de projeter automatiquement ces patrons, sans vérification manuelle, ce qui suppose de concevoir des patrons qui permettent le plus efficacement de repérer la construction visée. Afin de limiter le bruit, nous avons donc imposé deux conditions à la validation d'un couple extrait par les patrons :

- au moins un des éléments du couple extrait doit figurer dans la liste des 7278 mots qui ont au moins un antonyme dans le *Trésor de la langue française* ;
- X et Y doivent appartenir à la même classe grammaticale – Fellbaum (1995) a en effet remarqué que la cooccurrence de mots de sens opposés appartenant à des classes grammaticales différentes ne donnait jamais

lieu à des structures récurrentes.

Nous verrons que ces dernières sont toutefois insuffisantes pour filtrer la totalité du bruit généré par les patrons. Ainsi, au vu des premiers résultats fournis, nous nous sommes aperçus que les patrons *de X à Y* et *X comme Y* généraient une quantité particulièrement importante de bruit :

(8) Avant *de* **chercher** *à* **comprendre** le fonctionnement des syllogismes, il faut prendre garde à un point des plus importants [...].

(9) Confirmant cette analyse, la voiture compressée de 1997 intitulée Skin crime 2 (Givenchy 601) se réfère de manière formelle, par son titre, à une collection *de* **rouges** *à* **lèvres** de la marque Givenchy.

(10) Ensuite, en combat, la touche triangle sert de raccourci pour effectuer certaines **actions** *comme* les **attaques** en coopération, la fusion ou bien encore une invocation suivant l'activation de la compétence concerné dans le menu (et sous certaines conditions).

(11) Il met en valeur les monuments par des jeux de lumière, projette des images gigantesques sur les façades des immeubles, utilise les toits des **bâtiments** *comme* **base** de lancement de feux d'artifices, etc.

De ce fait, nous avons écarté ces deux patrons de notre jeu.

Les exemples ci-dessus montrent que la composante syntaxique joue un rôle de premier ordre dans la qualité des résultats obtenus. Quand c'était possible, nous avons exploité les liens de dépendance entre X, Y et les éléments du patron. Même si, comme le signale Murphy (2006), les constructions antonymiques ne respectent pas forcément le cadre des constituants syntaxiques, ce niveau d'information s'est avéré précieux pour contraindre l'extraction. Par exemple l'extraction du patron *X ou Y* consiste simplement à suivre les liens de coordination posés par Syntex. À l'inverse, le traitement des cas où les liens syntaxiques ne sont pas reconnus (c'est le cas du patron *ni X ni Y*, par exemple) nécessite la construction d'expressions régulières spécifiques dont il nous a fallu régler les paramètres. Nous avons par exemple fait le choix d'opérer exclusivement au niveau du mot et d'ignorer les constituants que sont les syntagmes, ce qui donne lieu à des erreurs d'extraction. Dans les exemples suivants, nous avons noté en gras les couples de mots extraits, en italiques le patron, entre crochets le syntagme complet qu'il faudrait identifier pour repérer correctement la relation d'opposition :

(12) [Homme de **cabinet**] *plutôt que* **guerrier**, Clarke fut un administrateur habile et intègre [...].

(13) Si le défunt mari **régnait** *ou* [**portait** un titre], on parlera d'impé-

ratrice douairière [...].

(14) Des instruments s'effacent, d'autres **apparaissent** *ou* [**prennent** leur forme définitive] [...].

Une tentative de restitution des syntagmes dans les phrases du corpus s'est en effet avérée infructueuse tant la nature cohésive des éléments constituant le syntagme est changeante : il serait par exemple pertinent d'extraire les noms *palais* et *château* des segments *palais de marbre* et *château de pierre*, alors que *palais de justice* ou *château d'eau* doivent conserver leur unicité. La difficulté à contrôler la taille des SN et SV à considérer nous a conduit, dans cette première étude, à nous limiter à l'extraction des paires d'adjectifs.

La projection des patrons nous a permis d'extraire 33 698 couples d'adjectifs (soit environ 33 % du nombre de couples extraits toutes catégories confondues). À titre d'illustration, nous rapportons ci-dessous quelques-uns des résultats obtenus :

(15) La promenade Malecón autour du port attire beaucoup de touristes, *plus* **nationaux** *qu'***internationaux**.

(16) La violence à l'école : phénomène **central** *ou* **marginal** ?

(17) Au cours de cet examen, Gon va rencontrer différents personnages, *aussi bien* **amis** *qu'***ennemis** [...].

(18) Selon les époques et les auteurs, cette divinité pouvait être *soit* **bienveillante** *soit* **malfaisante**.

(19) Le méso-climat spécifique induit des vendanges tardives qui donnent aux vins de Bellet un caractère *plus* **septentrional** *que* **méridional**.

(20) Dieu ne contient aucune différence, *ni* **temporelle** *ni* **spatiale**.

(21) Toujours pensée comme porteuse de valeur ou comme valeur incarnée, la notion d'œuvre se révèle aujourd'hui *moins* **descriptive** *que* **normative** [...].

(22) Ils découvrirent alors que les rayons X avaient la propriété d'ioniser l'air, puisqu'ils purent montrer que cela produisait de grandes quantités de particules chargées, *autant* **positives** *que* **négatives** [...].

Parmi l'ensemble des couples extraits, seuls 35 % – 11 751 – sont uniques (l'ordre d'apparition des éléments X et Y dans le patron a été ignoré pour ce calcul). On peut en effet voir au tableau 6.4, dans lequel nous avons reporté les dix couples les plus fréquemment extraits, que certaines paires apparaissent plusieurs centaines de fois.

fréquence	couple
581	*étranger/français*
562	*égal/supérieur*
492	*égal/inférieur*
220	*partiel/total*
215	*direct/indirect*
209	*bon/mauvais*
204	*négatif/positif*
182	*international/national*
182	*féminin/masculin*
176	*civil/militaire*

TAB. 6.4 – Les dix couples les plus fréquemment extraits par les patrons.

	toutes catégories confondues	couples adjectivaux seulement
X ou Y	75 389 (73,94 %)	28 318 (84,03 %)
entre X et Y	18 750 (18,39 %)	1992 (5,91 %)
à la fois X et Y	2350 (2,30 %)	1365 (4,05 %)
ni X ni Y	1746 (1,71 %)	628 (1,86 %)
X plutôt que Y	973 (0,95 %)	197 (0,58 %)
aussi bien X que Y	840 (0,82 %)	319 (0,95 %)
plus X que Y	671 (0,66 %)	543 (1,61 %)
soit X soit Y	646 (0,64 %)	195 (0,58 %)
autant X que Y	428 (0,42 %)	86 (0,26 %)
plutôt X que Y	90 (0,09 %)	19 (0,06 %)
moins X que Y	81 (0,08 %)	36 (0,11 %)
total	101 964	33 698

TAB. 6.5 – Nombre de couples extraits par les patrons.

Nous avons rapporté au tableau 6.5 le détail du nombre de couples extraits par chaque patron. On constate une prédominance du patron *X ou Y* (74 % des couples rapportés apparaissent dans cette structure, 84 % dans le cas des couples d'adjectifs). Parmi l'ensemble des patrons dont nous disposons, celui-là est celui qui impose le moins de contrainte sur les propriétés sémantiques des éléments qu'il met en opposition (contrairement à *plus X que Y*, par exemple, qui n'accepte, théoriquement, que des éléments X et Y qui ont la propriété d'être gradables). De ce fait, nous avons décidé de nous limiter à un sous-ensemble de couples, à savoir ceux qui sont reliés par au moins deux patrons de type distinct (en général, *X ou Y* et un autre patron), soit 907 couples. Cette démarche fait écho à celle de Jones *et al.* (2007), qui émet l'hypothèse que la diversité des constructions dans lesquelles apparaissent deux mots opposés est un meilleur indicateur de leur canonicité que leur fréquence.

On peut voir que certains couples qui seraient difficilement identifiés comme des antonymes émergent malgré la rigueur du filtrage que nous avons imposé. Après avoir analysé un échantillon de 100 couples, nous pouvons estimer leur proportion à 25 % (ce chiffre est toutefois à prendre avec précaution tant la nature antonymique de deux mots est dépendante de leurs contextes d'apparition). Par exemple, *translucide* et *transparent* semblent clairement entretenir une relation de synonymie. Or, ils apparaissent dans les patrons antonymiques :

(23) Les jeunes poissons sont d'abord **transparents** *ou* **translucides**.

(24) Il peut être *soit* **translucide** *soit* **transparent**.

Dans l'exemple (23), l'ambiguïté du patron *X ou Y* nous permet de penser que dans ce cas, la négation se fait au niveau métalinguistique, qu'on a affaire à une reformulation. L'exemple (24) est plus difficile à interpréter. On peut également considérer comme non antonymique le couple *social*/*littéraire* :

(25) La pénétration de la littérature yiddish sous sa forme moderne qui a suivi Abramovitsh, démontre l'importance de donner une voix aux aspirations juives *aussi bien* **sociales** *que* **littéraires**.

(26) Afin de mettre à l'épreuve sa théorie, McGrady recruta plusieurs collaborateurs au journal (cinq femmes et dix-neuf hommes, soit 24 auteurs en tout) pour qu'ils élaborent un roman au caractère sexuel explicite, sans aucune valeur **littéraire** *ou* **sociale**, quelle qu'elle soit.

Dans ce cas, on peut considérer qu'on a affaire à un cas d'incompatibilité

(cf. typologie de Cruse (2004)) entre deux co-hyponymes d'une catégorie créée en discours. En effet, Jones a montré que les patrons qu'il a repérés avaient la propriété de pouvoir renforcer le caractère antonymique de deux mots déjà opposés mais aussi de générer de l'opposition entre deux mots sans lien sémantique *a priori*. Dans l'exemple ci-dessous, on peut considérer que l'opposition entre les deux adjectifs *administratif* et *humain* est induite, discursivement, par la construction contrastive. On connait de fait la capacité du discours d'instaurer des relations non lexicalisées (cf. section 4.1.2).

(27) Le terme, néanmoins, reflète des connotations *plus* **humaines** *qu'***administratives**.

(28) Mais cette fois, suite à une négligence **humaine** *ou* **administrative**, la vanne ne fut pas rouverte

À ce stade, nous disposons donc de deux méthodes – l'ADA et les patrons – qui captent l'une et l'autre des relations antonymiques. L'évaluation qui suit cherche à mesurer leur complémentarité, et plus particulièrement à déterminer si les couples repérés conjointement par les deux méthodes présentent un degré d'antonymie plus marqué. Sur les 907 couples retenus par la technique des patrons, 612 sont également des voisins (environ 67 %). L'hypothèse est que le fait que deux adjectifs soient proches distributionnellement permettrait de filtrer les cas de contrastes occasionnels produits en discours.

6.3 Évaluation des résultats

Nous proposons deux méthodes pour juger du caractère antonymique des paires d'adjectifs extraites : la première passe par l'utilisation d'une ressource de référence, la deuxième consiste à solliciter des locuteurs en leur demandant de classer les paires selon la relation sémantique qu'elles portent. Le recours à deux méthodes nous a semblé s'imposer au regard des descriptions de l'antonymie produites dans le champ de la linguistique cognitive (cf. section 6.1.2), qui insistent sur le caractère graduel de l'antonymie, dont une ressource dictionnairique ne rend pas compte.

6.3.1 Comparaison à une ressource de référence

Nous avons dans un premier temps vérifié si les couples extraits étaient présents dans le dictionnaire d'antonymes du CRISCO, consultable sur la plate-forme CNRTL[2] et sur le site du DES (les antonymes sont renvoyés

[2] `http://www.cnrtl.fr/antonymie/`

	antonymes	synonymes	absents	
voisins	198 (32,4 %)	26 (4,2 %)	388 (63,4 %)	612 (100 %)
non-voisins	81 (27,5 %)	17 (5,7 %)	197 (66,8 %)	295 (100 %)

TAB. 6.6 – Proportion des couples voisins et non voisins présents ou absents du dictionnaire.

conjointement aux synonymes pour une requête donnée). Le recours à une telle ressource a été motivé par le fait qu'elle nous permet de connaître la proportion de couples extraits par les patrons – et éventuellement par les voisins – qui entretiennent une relation antonymique que des lexicographes ont considérée comme suffisamment consensuelle pour apparaître dans un dictionnaire. De fait, on ne s'attend pas à ce que des relations d'opposition contextuelles comme *social/littéraire* ou *humain/administratif* soient captées par cette méthode. Ne disposant pas de version locale du dictionnaire, nous avons vérifié la présence des 907 couples manuellement, *via* les interfaces de consultation en ligne. Nous avons vu avec *transparent/translucide* que certains d'entre eux pouvaient relever de la synonymie. Nous avons ainsi choisi de prendre en compte le nombre de ceux qui étaient identifiés comme tels dans le DES.

Les résultats, rapportés au tableau 6.6, indiquent la répartition des couples voisins et non voisins parmi les trois catégories suivantes : les antonymes, les synonymes, et les couples absents du dictionnaire. Ils indiquent que 32,4 % des couples qui ont à la fois été extraits par les patrons et par l'ADA apparaissent dans le dictionnaire d'antonymes, alors que ce n'est le cas que de 27,5 % des couples qui ont seulement été extraits par les patrons. Le calcul du chi^2 montre cependant que cette différence n'est pas significative. Ces résultats semblent donc infirmer notre hypothèse selon laquelle le fait de croiser une approche par patrons et l'ADA pour capter les couples qui fonctionnent à la fois sur le plan syntagmatique et paradigmatique nous permet d'extraire le plus sûrement des paires d'antonymes. Ils sont toutefois à interpréter en gardant à l'esprit les limites inhérentes aux ressources dictionnairiques. Ainsi, comme on peut le voir parmi les exemples de couples rapportés au tableau 6.7, certaines des paires qui ne figurent pas dans le dictionnaire apparaissent pourtant comme clairement antonymiques (*acide/basique*, *nuisible/bénéfique*, *hétérosexuel/homosexuel*, etc.).

Une autre limite de la comparaison des couples extraits avec un dictionnaire est que ce dernier ne nous fournit aucune information sur le degré de canonicité des couples. Ainsi, on peut voir – toujours dans le tableau 6.7 – que des couples comme *actuel/potentiel* ou *pair/impair* apparaissent comme

	antonymes	synonymes	absents
voisins	*nuisible/utile* *pair/impair* *humide/sec* *inférieur/supérieur* *hétérogène/homogène*	*agressif/violent* *bon/juste* *gratuit/libre* *voisin/proche* *mécanique/physique*	*gazeux/liquide* *hétérosexuel/homosexuel* *acide/basique* *médiéval/moderne* *naturel/urbain*
non voisins	*gai/triste* *épais/liquide* *actuel/potentiel* *profane/religieux* *lisse/rugueux*	*rouge/roux* *facile/simple* *épais/large* *accidentel/fortuit* *facile*/simple	*nuisible/bénéfique* *aggressif/craintif* *distant/local* *physique/verbal* *possible/souhaitable*

TAB. 6.7 – Exemples de couples extraits par les patrons, voisins et non voisins, présents ou absents du dictionnaire.

également antonymes, alors qu'intuitivement le deuxième apparaît comme plus canonique. De la même façon, on remarque que *nuisible/utile* est recensé dans le dictionnaire alors que *nuisible/bénéfique* ne l'est pas, ce qui, encore une fois, pourrait paraître contre-intuitif.

De ce fait, nous avons choisi de compléter l'approche basée sur l'utilisation d'un dictionnaire par une démarche qui a consisté à demander à des locuteurs de classer les différentes paires extraites par les patrons. Cette dernière approche, inspirée des travaux en psycholinguistique évoqués à la section 6.1, nous apparaît en effet plus à même de nous apporter des réponses à la question de savoir si le phénomène de canonicité se manifeste particulièrement sur les couples qui ont un fonctionnement à la fois syntagmatique et paradigmatique.

6.3.2 Questionnaires

Ce type d'approche a largement été utilisé dans le cadre de l'étude de l'antonymie, que ce soit sous la forme de tests d'association de mots (Deese, 1964) ou de diverses tâches consistant notamment à classer des paires d'antonymes selon le degré d'opposition ressenti (Herrmann *et al.*, 1986). Le protocole de notre expérience se rapproche du type d'études menées par Herrmann et collègues, dans le sens où l'on présente aux participants des couples de mots à évaluer. Nous avons extrait aléatoirement trois jeux de 100 paires parmi les 907 qui ont été extraites par les patrons. La seule contrainte a porté sur le fait que chacun de ces jeux devait être composé pour moitié de couples apparaissant parmi les voisins. Chacun des trois sous-ensembles a été pré-

			opposition					
nom(s) modifié(s)	adj. 1	adj. 2	forte	partielle	syno.	autre rel.	pas de rel.	n. s. p.
saut évolutif, évolution	graduel	important					✓	
caractère	improbable	probable	✓					
espace vide, univers	homogène	isotrope			✓			
affaire, réseau, modèle...	domestique	professionnel		✓				
lampe, béret, fruit...	rouge	vert				✓		
palais, gymnase, auteur...	grec	romain		✓				
opposition, information...	public	secret	✓					
non-métal	ductile	malléable			✓			
activité, débat, recherche...	artistique	scientifique		✓				
épidémie, phénomène	intentionnel	naturel		✓				
anneau	factoriel	principal						✓

TAB. 6.8 – Extrait d'un des formulaires soumis aux participants de notre étude.

senté à deux locuteurs – 1 étudiant de master, 4 doctorants et un maître de conférences en sciences du langage – à qui il était demandé de classer chaque paire selon la relation sémantique unissant ses deux membres. La consigne était la suivante :

> Est-ce que, selon vous, les adjectifs 1 et 2, qui modifient le ou les nom(s) (ou groupes nominaux) de la colonne 1 :
> – entretiennent une relation d'opposition (ressentie comme "forte" ou "partielle") ?
> – sont plus ou moins synonymes ?
> – sont liés par un autre type de relation ?
> – n'ont rien à voir l'un avec l'autre ?

Nous avons reproduit au tableau 6.8 un extrait des formulaires que nous avons soumis aux participants (les réponses qui y figurent sont authentiques). Afin d'assister les sujets dans leurs choix, chaque paire était présentée accompagnée d'un ou plusieurs éléments de contexte, à savoir des exemples de noms modifiés par les adjectifs en question (*lampe*, *béret*, *fruit* pour la paire *rouge/vert*). Les catégories qui ont été proposées aux participants étaient les suivantes : `opposition forte`, `opposition partielle`, `synonymie`, `autre relation`, `pas de relation` et `ne sais pas`. En divisant deux types d'opposition – *forte* vs *faible* –, nous cherchons à distinguer les couples d'antonymes canoniques, pour lesquels l'opposition est manifeste – *probable/improbable* et *secret/public* dans notre extrait –, de ceux qui sont perçus comme des mots dont les sens sont opposés sans pour autant provoquer de réaction aussi franche que pour les antonymes canoniques (leur opposition est plus contextuelle, moins immédiatement perceptible) – *domestique/professionnel*, *artistique/scientifique*. Ces dernières paires de mots

sont alors dites *non canoniques*. Le consensus autour de leur statut étant moins clair parmi les locuteurs, nous avons proposé dans notre expérience d'autres catégories dans lesquelles classer ces couples, notamment la catégorie `synonymie` (*homogène/isotrope*, *ductile/malléable*). La catégorie `autre relation` accueille principalement des co-hyponymes comme, dans notre extrait, *rouge/vert*. La catégorie `pas de relation` est destinée aux paires de mots ne présentant pas de relation sémantique identifiable malgré la présence de contextes communs. Le rapprochement entre les deux adjectifs peut alors être perçu comme accidentel, comme pour le couple *graduel/important* de notre extrait. Enfin, la catégorie `ne sais pas` (`n. s. p.`) permet de classer les paires pour lesquelles les sujets estiment que leur compétence ne leur permet pas d'établir de jugement, ce qui a été le cas pour le participant qui a produit les réponses de notre extrait et qui ne s'est pas prononcé sur la nature de la relation entre *principal* et *factoriel*. Le formulaire contenait une dernière colonne dans laquelle il était proposé aux participants de faire un commentaire sur les raisons qui les ont poussés à annoter un couple d'une certaine façon. Cette possibilité n'a toutefois été saisie que par un annotateur particulièrement coopératif qui a laissé quelques commentaires qui illustrent la difficulté de la tâche :

- pour le couple *rouge/vert* avec les contextes *lampe*, *béret*, *fruit*, qui a été annoté comme relevant d'une opposition partielle :

  > "Pour béret je sais pas combien il en existe de différents mais si c'est que deux c'est opposition forte."

- pour le couple *matériel/moral* avec les contextes *destruction*, *preuve*, *difficulté*, qui a été annoté *ne sais pas* :

  > "Je sais pas, totale dans langage juridique, partiel dans un autre domaine ?"

Le fait de choisir trois paires de deux locuteurs pour notre expérience a été en partie motivé par l'intérêt qu'il y a à observer l'accord inter-annotateurs : nous avons en effet pu constater que lorsque l'opposition n'était pas immédiatement ressentie comme canonique, l'attribution d'une catégorie à une paire était plus difficile. Ainsi, sur les trois jeux, le kappa de Cohen moyen pour la relation d'opposition forte est de 0,64 alors qu'il est de 0,27 pour l'opposition faible.

Les résultats obtenus à l'aide des questionnaires ont été rapportés au tableau 6.9. On peut par exemple voir que le locuteur 1 a annoté un total de 30 couples sur 100 comme relevant d'une opposition forte, parmi lesquels 14 apparaissent dans les voisins. Ainsi, alors que nous nous attendions à observer une majorité des couples annotés `opposition forte` aux lignes `+ voisin`, on s'aperçoit que 5 des 6 locuteurs ont annoté davantage de couples de non-

			opposition					
			forte	partielle	syno	autre	pas de rel.	n. s. p.
jeu 1	loc. 1	+ voisin	14	22	7	4	3	0
		- voisin	16	18	10	2	3	1
	loc. 2	+ voisin	16	14	1	1	16	2
		- voisin	17	9	10	0	9	5
jeu 2	loc. 3	+ voisin	16	9	2	15	7	1
		- voisin	17	9	2	3	13	6
	loc. 4	+ voisin	24	7	3	12	4	0
		- voisin	26	8	2	8	5	1
jeu 3	loc. 5	+ voisin	18	10	5	8	8	1
		- voisin	16	9	5	8	8	4
	loc. 6	+ voisin	18	13	6	12	1	0
		- voisin	23	9	8	6	4	0
	moyenne	+ voisin	17,7	12,5	4	8,7	6,5	0,6
		- voisin	19,2	10,3	6,2	4,5	7	2,8

TAB. 6.9 – Résultats obtenus après le dépouillement des questionnaires.

voisins comme relevant d'oppositions fortes. On voit qu'en moyenne, 17,7 couples de voisins sur 50 ont été annotés `opposition forte` contre 19,2 couples de non-voisins. En revanche, le jugement d'opposition partielle est plus favorable aux couples de voisins : seul le locuteur 4 a annoté davantage de couples de non-voisins comme des oppositions partielles. Le nombre moyen de couples annotés `opposition partielle` est cette fois supérieur pour les couples de voisins.

Globalement, les résultats ne montrent pas de différence de jugement significative chez les locuteurs en ce qui concerne l'opposition forte. Cela corrobore le constat que nous avions fait lors de l'utilisation du dictionnaire d'antonymes : le fait que deux adjectifs trouvés dans des patrons antonymiques soient des voisins distributionnels ne semble pas influer sur leurs chances de porter une relation d'antonymie canonique.

6.3.3 Analyse

Les résultats obtenus semblent contre-intuitifs : ils indiquent que la combinaison du critère de substituabilité au critère de cooccurrence ne permet pas de capter des couples qui sont perçus plus antonymiques que ceux qui ne sont pas substituables. Nous pouvons à ce stade proposer plusieurs éléments d'explication à ce constat (les deux premiers tiennent plutôt aux caractéristiques de l'ADA et du corpus, le dernier fournit des éléments d'observation

intéressants sur la relation d'antonymie) :

- les mots opposés sont trop rares pour être captés par l'ADA. C'est par exemple le cas des mots du couple *dioïque/monoïque*, qui apparaissent respectivement à 50 et 62 occurrences dans le corpus. Cette remarque vaut également pour les couples dont l'un des membres a une fréquence trop faible pour que le score de la paire ne dépasse ce seuil. N.B. : chronologiquement, ce travail sur l'antonymie a été le premier que nous avons mené dans le cadre de notre thèse. De ce fait, ce n'est qu'à partir de cette étude que nous avons commencé à prendre des précautions méthodologiques en écartant les couples de productivités trop inégales (cf. préambule méthodologique p. 119) ;
- deux antonymes peuvent s'opposer lorsqu'ils modifient un ensemble très limité de noms. C'est notamment le cas de *annuel* et *vivace*, qui ne s'opposent que lorsqu'ils portent sur les noms *plante* ou *espèce*, ou de *ras* et *long*, qui ne modifient que *poil*. Bien que ces couples soient extraits par les patrons, ils partagent un nombre trop limité de contextes pour être détectés comme des voisins distributionnels ;
- le dernier cas de figure nous place au cœur de l'hypothèse de Murphy, Jones et leurs collègues, à savoir le comportement syntagmatique de la relation d'antonymie : le fait que certaines paires d'antonymes ne soient pas repérées par l'ADA permet en effet de dégager un sous-ensemble de couples qui privilégient la relation de cooccurrence à la relation de substituabilité. On a alors affaire à des couples qui fonctionnent sur un mode quasi locutionnel. C'est en particulier le cas de paires comme *officiel/officieux*, *mélioratif/péjoratif* ou *vertueux/vicieux*. La combinaison des deux méthodes d'analyse fournit alors des éléments d'observation qui permettent d'affiner la description de la relation d'antonymie, en distinguant des paires pour lesquelles le double plan paradigmatique et syntagmatique fonctionne à plein, et d'autres pour lesquelles c'est le plan syntagmatique qui prime.

6.4 Conclusion

Notre objectif était d'apporter de nouveaux éléments d'expérimentation issus de techniques de TAL pour tester l'hypothèse d'un double fonctionnement paradigmatique et syntagmatique de l'antonymie, formulée par une série de travaux récents en linguistique cognitive et linguistique de corpus. Les résultats que nous avons obtenus semblent aller contre cette hypothèse en montrant que des paires de mots trouvées dans des contextes contrastifs n'ont pas plus de chance d'être perçues comme antonymiques lorsqu'elles ap-

partiennent à une même classe distributionnelle – contrairement à l'intuition qui consisterait à penser que le test de substituabilité fournirait un filtre efficace pour séparer les vrais antonymes des paires de mots qui entretiennent une relation d'opposition très éphémère suggérée par des contextes discursifs particuliers.

Nous avons pu constater, dans le tableau 6.7, que cette méthode met par ailleurs au jour des paires d'antonymes qui ne sont pas voisins distributionnels, ce qui permet de dégager des cas prototypiques d'association purement syntagmatique, qui tranchent avec l'hypothèse classique d'un fonctionnement paradigmatique.

Ces premiers résultats ouvrent des pistes pour l'étude du continuum entre fonctionnement paradigmatique et syntagmatique. Une autre perspective consisterait à vérifier l'hypothèse d'un comportement particulier de l'antonymie au sein des relations lexicales, et à s'intéresser cette fois à l'articulation entre les dimension syntagmatique et paradigmatique pour les relations de synonymie et d'hyperonymie.

Chapitre 7

Observer la substituabilité des hypo/hyperonymes dans une base distributionnelle

Sommaire

Après avoir abordé, dans les deux chapitres précédents, la question des modalités du repérage des synonymes et des antonymes dans les bases distributionnelles, nous nous penchons à présent sur le cas des hypo/hyperonymes. La relation d'hyperonymie joue un rôle primordial dans la structuration des ontologies et des réseaux *à la WordNet*. Son repérage automatique représente donc un enjeu de taille. L'approche par patrons initiée par Hearst (1992) est encore aujourd'hui la méthode la plus répandue pour extraire les couples d'hypo/hyperonymes à partir de textes. Dans la section 4.3.3.2, nous avons mesuré le potentiel de l'analyse distributionnelle automatique (ADA) à capter la relation d'hyperonymie : le croisement des voisins de Wikipédia (VDW) avec JeuxDeMots (JDM) a montré que moins de la moitié des relations IS A et HYPO contenues dans JDM étaient extraites par l'ADA. L'exploitation de cette méthode est toutefois gênée par le fait que, d'une part, les relations d'hyperonymie ne sont pas identifiées dans la masse des voisins, et, d'autre part, par le fait que les modalités qui régissent l'extraction des voisins hypo/hyperonymes restent mal connues. On ignore par exemple les raisons qui font que la plupart des couples IS A et HYPO ne sont pas extraits par l'ADA.

Dans l'étude que nous menons ici, nous abordons ainsi la question de savoir pourquoi certains couples d'hypo/hyperonymes ont tendance à partager les mêmes contextes d'apparition – et sont donc captés par l'ADA – alors que ce n'est pas le cas pour d'autres. Nous avons choisi de filtrer les couples de voisins porteurs d'une relation d'hyperonymie en croisant nos trois bases distributionnelles avec JDM. Cela nous permet de disposer de deux ensembles de couples d'hypo/hyperonymes, à savoir ceux qui ont été extraits par l'ADA et ceux qui ne l'ont pas été. Notre approche consiste à extraire et à mettre en parallèle trois ensembles d'hyperonymes :

1. ceux dont les hyponymes ont été captés en masse dans les trois bases de voisins ;
2. ceux pour lesquels, au contraire, les hyponymes n'ont été que peu extraits par l'ADA, et ce dans les trois bases de voisins ;
3. ceux pour lesquels la proportion d'hyponymes extraits varie d'une base de voisins à l'autre.

Ces trois ensembles de mots nous offrent ainsi la possibilité d'étudier des cas de décalage distributionnel entre un hyperonyme et ses hyponymes ainsi

que la stabilité de ce phénomène en fonction de la base de voisins considérée. Le deuxième de ces trois cas, celui des hyperonymes dont les hyponymes ne sont pratiquement pas repérés dans les trois bases de voisins, est celui qui retiendra le plus notre attention : il va en effet à l'encontre de l'intuition selon laquelle, selon le principe de l'héritage des propriétés sémantiques d'un hyperonyme vers ses hyponymes, les hyponymes devraient partager les contextes d'apparition de leurs hyperonymes. Nous mettons au jour plusieurs phénomènes qui expliquent ce décalage distributionnel entre hyponymes et hyperonymes.

Dans un premier temps, à la section 7.1, nous présentons les propriétés sémantico-logiques de la relation d'hyperonymie et, en particulier, la notion d'héritage, qui nous permet d'aborder le sujet de la substituabilité des hypo/hyperonymes. Nous présentons ensuite, à la section 7.2, la démarche de comparaison entre les voisins et JDM que nous avons adoptée ici. La section 7.3 décrit les différents filtrages que nous avons mis en place pour extraire les couples les plus représentatifs des phénomènes que nous cherchons à analyser. Les analyses que nous avons faites de ces données sont rapportées à la section 7.4. Celle-ci est divisée en trois sections qui portent chacune sur les trois types d'hyperonymes que nous avons évoqués plus haut.

7.1 Propriétés de la relation d'hyperonymie

Nous entamons cette étude en rappelant quelques propriétés sémantico-logiques de la relation d'hyperonymie. Cela nous amène à aborder la notion d'héritage, sur laquelle s'appuie le principe de substituabilité des couples d'hypo/hyperonymes. Nous évoquons pour finir quelques-unes des limites auxquelles se retrouve confronté le critère de substituabilité.

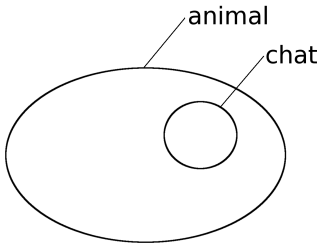
7.1.1 La notion d'inclusion

La relation d'hyperonymie se définit comme une relation d'inclusion hiérarchique entre deux éléments A et B. Si A inclut B, A est l'hyperonyme de B. Contrairement aux relations symétriques que sont la synonymie et l'antonymie, l'hyperonymie est une relation orientée. De ce fait, la relation entre B et A n'est pas la même qu'entre A et B : on parle alors d'hyponymie. Afin d'éviter les lourdeurs, nous utilisons par la suite le terme *hyperonymie* pour désigner à la fois la relation entre A et B et sa réciproque.

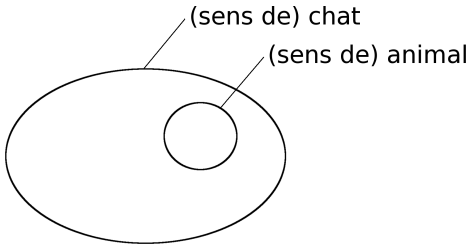
Telle que nous l'avons définie, la relation d'hyperonymie est ambigüe. En effet, l'orientation de la relation d'inclusion varie selon qu'elle est considérée

du point de vue extensionnel ou intensionnel (Kleiber et Tamba, 1990) :

– du point de vue extensionnel, A et B renvoient à des référents du monde : le chat est un animal, donc la classe des chats est comprise dans la classe des animaux (tout ce qui est un chat est un animal). C'est cette définition de la relation qui est adoptée dans les ontologies et réseaux lexicaux comme WordNet. Le rapport d'inclusion peut être schématisé de la façon suivante :



– du point de vue intensionnel, A et B renvoient aux sens des référents *chat* et *animal*. Le sens de *chat* englobe celui de *animal*, auquel viennent s'ajouter des sèmes spécifiques comme /possède quatre pattes/ ou /miaule/. Le rapport d'inclusion est alors inversé :



Nous verrons que dans notre étude, il nous faudra prendre en compte ces deux acceptions de la relation d'inclusion. Les couples de JDM que nous utilisons ont été produits par des locuteurs pour qui la relation d'inclusion est envisagée du point de vue extensionnel, qui apparaît beaucoup plus intuitif que le point de vue intensionnel. Ce dernier nous permet toutefois d'étudier le caractère substituable des hypo/hyperonymes à travers la notion d'héritage, que nous abordons à la section 7.1.3.

7.1.2 Hypo-hyperonymie et définition lexicographique

La relation d'hyperonymie est un des éléments constitutifs de la définition lexicographique traditionnelle. Dans le TLF, le mot *chat*, dans son sens courant, se définit de la façon suivante :

> Petit animal domestique carnassier, à pelage de couleur variée souvent noir ou gris, se nourrissant de souris, de petites proies, et de la nourriture servie par ses maîtres.

On remarque que *chat* est en premier lieu défini par son hyperonyme *animal*, puis par un ensemble de propriétés qui lui sont spécifiques ("à pelage de couleur variée", "se nourrissant de souris"). On peut faire le parallèle entre cette utilisation de l'hyperonymie et les notions de *genus* et de *differentiae* de la définition aristotélicienne[1], que nous avons évoquées à la section 5.4.4.1 :

- le *genus* renvoie à la classe à laquelle appartient le mot ;
- les *differentiae* correspondent aux traits qui vont permettre de distinguer l'élément à définir des autres éléments appartenant à la même classe.

La notion de *genus* se rapproche de celle d'hyperonyme dans le sens où il s'agit d'un concept superordonné qui englobe le sens du concept subordonné (dans le cas d'une définition intensionnelle de l'hyperonymie). Le sens d'un mot correspond donc à la somme des traits du concept superordonné – qui ne sont pas explicités dans la définition – et des *differentiae*.

7.1.3 Hyperonymie et substituabilité

L'héritage des propriétés d'un hyperonyme par ses hyponymes entraîne ainsi un certain degré de similarité sémantique entre eux. Cette similarité se traduit dans une certaine mesure au niveau de leurs distributions, puisque, de fait, on peut s'attendre à ce que les propriétés qui s'appliquent à *animal* s'appliquent également à *chat*. Par exemple, dans le corpus Wikipédia, les deux mots apparaissent en position sujet de *vivre*, *manger* et *posséder*, ou encore en position objet de *nourrir*, *aimer* ou *représenter*. Cette corrélation est notamment évoquée par Murphy (2003, p. 217) :

> Grammatically, selectional restrictions on (for example) the object of a verb can be phrased in terms of a hyperonym, and all hyponyms of that word are then also selected as potential objects (Resnik, 1993). For instance, *drink* selects for *beverage* and all its hyponyms (*water*, *beer*, *juice*, etc.).

[1] Voir Jackson (2002) pour plus de détails sur ces deux notions ou encore Brousseau et Roberge (2000), chez qui elles sont présentes sous le nom de *classificateurs* et *distingueurs*.

couples	exemples de contextes
palais/*bâtiment*	*loger*_DANS, *fronton*_DE, *façade*_DE
colère/*émotion*	*ressentir*_OBJ, *susciter*_OBJ, *éprouver*_OBJ
*marcher*_SUJ/*se déplacer*_SUJ	*forces*, *troupe*, *joueur*
jaune/*coloré*	*pigment*_MOD, *plumage*_MOD, *plastique*_MOD

TAB. 7.1 – Exemples de couples d'hypo/hyperonymes captés par l'analyse distributionnelle du corpus Wikipédia.

On s'attend donc à observer un chevauchement entre les contextes d'apparition de *animal* et de *chat*. Étant donné que, dans le cas des hypo/hyperonymes, seule une partie du sens est partagée, ce chevauchement est nécessairement partiel. S'il est suffisamment important, le couple sera capté par l'ADA, comme c'est le cas pour les couples présentés au tableau 7.1.

7.1.4 Limites du critère de substituabilité

La corrélation entre héritage d'un noyau de sens et héritage des contextes d'apparition que nous venons d'évoquer rencontre toutefois ses limites lorsque l'on considère les hypo/hyperonymes en contexte. Nous montrons dans cette section que l'intuition ainsi que les données en corpus vont à l'encontre de cette conception selon laquelle le principe de substituabilité des couples d'hypo/hyperonymes est entièrement régi par le principe d'héritage.

7.1.4.1 Une relation asymétrique

Dans le corpus Wikipédia, les mots *animal* et *chat* apparaissent respectivement dans 790 et 231 contextes différents[2]. Ils en partagent 115. Cela revient à dire qu'environ 50 % des contextes de *chat* sont communs aux deux mots, alors que ce n'est le cas que d'environ 15 % des contextes de *animal*. On peut attribuer ce phénomène au fait que l'hyperonyme relève – par définition – d'un degré d'abstraction supérieur à celui de ses hyponymes, ce qui entraînerait une diversité plus élevée dans sa distribution. Cela est confirmé par une mesure de la productivité moyenne des hyperonymes et des hyponymes de JDM dans le corpus Wikipédia : les hyperonymes apparaissent en moyenne dans 425 contextes différents contre 260 pour les hyponymes. Ainsi, nous pouvons définir la substituabilité des hypo/hyperonymes comme étant un phénomène fondamentalement asymétrique.

[2]Ces contextes apparaissent au minimum à 5 reprises avec *animal* ou *chat* dans le corpus (cf. section 3.2.2).

Ces chiffres laissent entrevoir les problèmes que pose le fait d'utiliser le critère de substituabilité pour définir la relation d'hyperonymie. Parmi les 116 contextes de *chat* dans lesquels *animal* n'apparaît pas, on peut certes trouver des cas comme *sauter*_SUJ ou *dormir*_SUJ qui seraient potentiellement compatibles avec *animal*. En revanche, *animal* ne pourra pas être modifié par *siamois* ou *haret*. Réciproquement, il est très peu probable – mis à part dans un corpus de textes imaginaires – que *chat* soit modifié par des adjectifs comme *ovipare*, *cavernicole*, ou ait pour complément *de boucherie*, contrairement à *animal*. Cela est dû au fait qu'un hyponyme et son hyperonyme relèvent, par définition, de degrés d'abstraction différents, et que certains contextes ne tolèrent pas le changement de degré d'abstraction.

7.1.4.2 Substituabilité et intuition

Dans l'exemple (1), on peut voir que l'emploi de certains hyperonymes de *chat* paraît peu intuitif dans la phrase *Le chat de Camille ronronne tendrement sur le fauteuil* :

(1)
$$\left.\begin{array}{c} \textit{Le chat} \\ \wedge \\ ?\ \textit{Le félin} \\ \wedge \\ ?\ \textit{Le mammifère} \\ \wedge \\ ?\ \textit{L'animal} \\ \wedge \\ ?\ \textit{L'être vivant} \\ \wedge \\ ?\ \textit{La chose} \end{array}\right\} \text{de Camille ronronne tendrement sur le fauteuil.}$$

Par exemple, l'emploi des hyperonymes *mammifère* ou *félin* paraît ici incongru. Cela peut s'expliquer par le fait que les mots *mammifère* et *félin* renvoient à des individus qui se définissent par leur appartenance à une classe taxonomique particulière. Or, cette information n'est pas pertinente au vu du contexte auquel on a affaire (elle le serait dans un texte relevant du domaine zoologique, ce qui n'est clairement pas le cas ici). De plus, ces deux noms relèvent d'un niveau d'abstraction tel que leur instanciation dans un contexte comme celui qui est présenté en (1) crée une sensation de décalage. En effet, les mots *mammifère* et *félin* semblent plus enclins à des emplois génériques, en particulier dans des constructions anaphoriques, comme l'illustre le passage suivant extrait de l'article *Josephoartigasia monesi* du corpus Wikipédia :

Josephoartigasia monesi est un rongeur dinomyidé éteint décou-

> vert à l'état fossile en Uruguay. **Ce mammifère** serait, selon ses auteurs, le plus grand et le plus lourd des rongeurs de tous les temps, supplantant ainsi le *Phoberomys pattersoni*, appartenant à la même famille.

De la même façon, l'emploi d'un hyperonyme de haut niveau comme *être vivant* ou *chose* semble également curieux, dans le sens où l'on se place à un niveau d'abstraction inutilement élevé.

Le niveau d'abstraction dans l'utilisation d'un nom donné correspond à un choix de la part du locuteur/scripteur. Il peut être motivé du point de vue discursif. Ainsi, l'emploi de *siamois* à la place de *chat* dans l'exemple (1) apporterait à la phrase des informations supplémentaires dont la pertinence s'évalue en fonction du contexte/cotexte. De tels exemples sont commentés chez Theissen (1997, p. 99) :

> L'emploi du niveau subordonné paraît ainsi légitimé par l'intérêt porté aux informations qui lui sont spécifiques avec l'idée que ces propriétés du référent encodées au niveau subordonné favorisent une compréhension plus précise de la situation.

Ainsi, le choix d'un mot en discours – pour paraphraser le titre de l'ouvrage de Theissen (1997) – est soumis à des contraintes contextuelles. Le critère de substituabilité se trouve alors mis à mal. Nous avons en effet vu que la relation entre hyperonymie et contraintes sélectionnelles était loin d'être aussi systématique que le laissait penser la citation de Murphy (2003) rapportée à la section 7.1.3. Car si *boire* sélectionne en effet aussi bien le mot *boisson* que ses hyponymes, on peut également trouver de nombreux contextes qui s'appliquent aux hyponymes de *boisson* mais qui, dans nos corpus, excluent le mot *boisson* lui-même. C'est notamment le cas des contextes *buveur*_DE, *couler*_SUJ ou encore *goutte*_DE, qui, dans les corpus Wikipédia, Le Monde et Frantext, sont incompatibles avec le mot *boisson* mais s'emploient avec ses hyponymes (*whisky*, *bière*, *lait*, etc.).

7.2 Croiser les voisins et les hypo/hyperonymes de JeuxDeMots

Dans la section précédente, nous avons montré que les propriétés logiques du principe de substituabilité se heurtaient à la réalité du discours : l'héritage des propriétés sémantiques d'un hyperonyme vers ses hyponymes ne semble pas se traduire de façon systématique sur le plan distributionnel. Le principe de substituabilité se situant au cœur même de la méthode

d'ADA, ce constat peut s'avérer problématique pour l'extraction des couples d'hypo/hyperonymes.

Pour vérifier dans quelle mesure ce phénomène entrave le repérage de la relation d'hyperonymie par l'ADA, nous croisons les hypo/hyperonymes de JDM et trois bases distributionnelles (les VDW, VDLM et VDF). Nous évoquons dans cette section les précautions à prendre lorsque l'on adopte une telle démarche. Nous présentons ensuite la méthode que nous avons adoptée pour croiser les voisins et JDM, puis nous montrons pourquoi, dans le cadre de cette étude, cette démarche est nécessairement comparative.

7.2.1 Une ressource de référence ?

Nous avons vu que le principal problème de l'ADA était que les relations qu'elle permet de mettre au jour ne sont pas sémantiquement typées. De ce fait, pour pouvoir étudier les propriétés distributionnelles des couples de voisins qui entretiennent une relation d'hyperonymie, il faut être en mesure d'identifier ces couples parmi les millions qui sont extraits par l'ADA. La démarche qui consiste à mobiliser un lexique externe contenant des couples d'hyperonymes possède l'avantage, d'une part, de nous permettre d'accéder aux couples de voisins qui sont des hypo/hyperonymes et, d'autre part, de nous fournir un ensemble de couples qui ne sont pas captés par l'ADA. Elle offre ainsi la possibilité de mener des études comparatives visant à mettre au jour les raisons qui font qu'un couple d'hypo/hyperonymes a été capté ou non.

L'inconvénient de cette méthode est qu'elle repose entièrement sur les données contenues dans le lexique. Dans notre cas, l'observation des manifestations de l'hyperonymie se fait alors sur des couples dont on assume qu'ils portent la relation d'hyperonymie. Or, comme nous le verrons avec les couples de méronymes que nous évoquons à la section 8.3.1, les couples d'hypo/hyperonymes que contient JDM relèvent parfois d'une hyperonymie *au sens large*. À titre d'exemple, on peut citer parmi les hyponymes recensés pour *argent*, les noms suivants : *billet*, *blé*, *bourse*, *chèque*, *dollar*, *économie*, *euro*, *franc*, *liquide*, *livre* et *pièce*. On voit que le problème naît ici du fait qu'une notion abstraite comme l'*argent* peut difficilement entrer dans une relation d'hyperonymie. Ainsi, le fait de placer *argent* et ses hyponymes *franc*, *euro* ou *dollar* dans un patron *X est (un type de) Y* fournit des résultats douteux :

(2) ? Un/l'euro est de l'argent.

(3) ? Un/l'euro est un type d'argent.

Les hyperonymes qui semblent les plus intuitifs dans ce cas seraient *devise*, *monnaie* ou *unité monétaire*.

On peut imputer la présence de ce type de couples dans le lexique à son mode de constitution. En effet, nous avons vu, à la section 4.3.3.1, que le réseau JeuxDeMots était enrichi par des internautes *via* un jeu en ligne qui consiste, pour un joueur, à fournir un maximum de réponses à une consigne donnée. Dans le cas de l'hyperonymie et de l'hyponymie, les consignes qui étaient présentées aux joueurs étaient les suivantes (Zampa et Lafourcade, 2011, p. 13) :

> Donner des GÉNÉRIQUES pour le terme qui suit (par exemple, *véhicule* pour *voiture*, *félin*, *animal* pour *chat*) ;
>
> Donner des SPÉCIFIQUES pour le terme qui suit (par exemple, *chat*, *chien*, *animal de compagnie*, etc. pour *animal* - ou encore *voiture*, *train*, *véhicule spatial*, etc. pour *véhicule*) ;

On note que les termes d'*hyponymie* et d'*hyperonymie* n'apparaissent pas, ce qui est dû au fait que ces mots relèvent d'un vocabulaire avec lequel les joueurs ne sont pas forcément familiers. Par ailleurs, le fait que les joueurs ne soient pas des spécialistes des relations lexicales peut expliquer un certain degré de *flottement* dans la définition de ce qu'est un terme specifique, à plus forte raison quand l'hyperonyme présenté est un mot comme *argent*. On voit alors que les hyponymes produits pourraient être davantage considérés comme des termes associés au sens large. Nous verrons à la section 8.1 que la confusion entre les relations peut même pousser les joueurs à produire des méronymes plutôt que des hyponymes.

7.2.2 Mesure du recouvrement entre les hypo/hyperonymes de JDM et les voisins

La version du réseau JeuxDeMots qui a été utilisée pour cette étude est celle du 28 septembre 2012. Elle compte 82 873 couples de mots portant la relation IS A – la relation d'hyperonymie – et 16 581 pour la relation HYPO.

Afin de permettre la comparaison avec les voisins, nous avons apporté à ces couples les modifications évoquées dans le préambule méthodologique. Nous avons ainsi :

1. écarté les paires dont au moins l'un des deux membres était absent des lexique des bases de voisins. Ainsi, la comparaison des couples d'hypo/hyperonymes et des voisins se fait sur la base du lexique qu'ils partagent (cf. 4.4.4) ;

	VDW	VDLM	VDF
+ voisins	5832	5044	2801
- voisins	9136	9820	7962
%	**39 %**	**33,9%**	**26 %**

TAB. 7.2 – Proportion d'hypo/hyperonymes de JDM dans les voisins de Wikipédia (VDW), Le Monde (VDLM) et Frantext (VDF).

2. symétrisé les couples d'hypo/hyperonymes : pour tout couple A/B où A est hyperonyme de B nous avons généré un couple B/A où B est hyponyme de A ;
3. effacé les doublons que cette symétrisation génère inévitablement (certains couples A/B de la relation d'hyperonymie figuraient également comme des couples B/A de la relation d'hyponymie).

Les résultats présentés au tableau 7.2 montrent qu'en moyenne, seulement un tiers – 33 % – des couples d'hypo/hyperonymes de JDM sont extraits par l'ADA. On peut observer une variation entre les bases distributionnelles, mais le recouvrement ne va pas au-delà de 39 %. Cela signifie que, pour chacune des ressources, la plupart des couples d'hypo/hyperonymes recensés dans JDM ne présentent pas un degré de similarité distributionnelle suffisant pour pouvoir être captés par l'ADA.

7.2.3 Une étude comparée

Sachant que les ressources générées à partir de corpus traduisent les fonctionnements à l'œuvre dans ces corpus, le fait de travailler avec une seule base distributionnelle entraîne donc un biais. Il devient en effet difficile – voire impossible – de savoir si un phénomène observé est particulier au type de corpus observé ou si l'on a affaire à un fonctionnement en langue, généralisable à d'autres types de données. Pour nous permettre de faire la distinction, nous avons mené notre étude sur trois bases de voisins distributionnels calculées sur des corpus appartenant à des genres différents : des articles d'encyclopédie (les VDW), des articles de journaux (les VDLM) et de la littérature (les VDF).

Les conséquences de la variation dans la distribution des mots du corpus peuvent s'observer dans les tableaux 7.3 et 7.4. Ils illustrent le cas d'un hyperonyme – *engin* – dont la proportion d'hyponymes extraits par l'ADA varie d'un corpus à l'autre, et le cas d'un hyperonyme – *émotion* – pour lequel cette proportion est relativement stable. Les symboles utilisés dans ces

tableaux sont les mêmes que ceux que nous avons employés dans le chapitre 5. Ils se lisent de la façon suivante :

- une coche verte – ✓ – signifie que l'hyponyme dans la première colonne est voisin avec le mot-cible (*argent* ou *émotion*) dans la base de voisins concernée ;
- une croix rouge – ✗ – indique que l'hyponyme est présent dans le lexique de la base de voisins mais n'est pas voisin avec le mot-cible, une absence de coche signifie que le mot n'est pas voisin avec le mot-cible ;
- une absence d'indication indique que l'hyponyme ne figure pas dans la base de voisins concernée (ce qui est souvent dû au filtrage que nous décrivons par la suite). Ces hyponymes n'ont pas été pris en compte dans le calcul du rappel.

La dernière ligne donne la proportion d'hyponymes voisins du mot-cible par rapport à l'ensemble de ses hyponymes (que nous appelons par la suite *rappel*).

La première chose que l'on remarque dans le tableau 7.3 est la différence de couverture lexicale des trois ressources. En effet, parmi les 12 hyponymes de *engin*, 9 sont absents des VDF. Cela peut aussi bien s'expliquer par la nature du corpus Frantext que par sa taille. Ce tableau montre également une différence dans la proportion d'hyponymes extraits : on voit que *engin* est relié à 87 % de ses hyponymes dans les VDW alors que ce n'est le cas que de 33 % dans les VDLM. Dans les VDF, aucun des trois hyponymes de *engin* présents dans le corpus n'ont été extraits.

Inversement, le tableau 7.4 montre une certaine stabilité entre les bases de voisins dans la proportion d'hyponymes extraits (le rappel – la proportion d'hyponymes captés par l'ADA – varie entre 0,79 et 0,91). On voit ainsi que l'hyperonyme *émotion* possède des affinités distributionnelles fortes avec ses hyponymes : il partage leurs contextes d'apparition dans les trois types de corpus considérés. La stabilité entre les corpus de la proportion d'hyponymes extraits est mesurée par la suite à l'aide de l'écart type. À ce stade, la question est donc de savoir pourquoi certains couples liés par une relation d'hyperonymie partagent les mêmes contextes d'apparition alors que d'autres ont des distributions qui divergent, et ce quel que soit le type de corpus observé.

7.3 Protocole

Nous partons ici du constat d'un décalage entre les couples de JDM et les voisins de Wikipédia afin d'étudier les conditions dans lesquelles deux hypo/hyperonymes sont ou ne sont pas substituables. Nous cherchons ainsi à expliciter les contraintes qui régissent la similarité distributionnelle entre

engin			
	VDW	VDLM	VDF
bus	✓	✗	✗
camion	✓	✓	
locomotive	✓		✗
métro	✗	✗	
ordinateur	✓	✓	
outil		✓	
pelle			✗
planeur	✓		
satellite	✓	✓	
train		✓	
voiture		✓	
wagon	✓	✗	
rappel	0,87	0,33	0

TAB. 7.3 – Hyponymes du nom *engin* dans les trois bases de voisins.

émotion			
	VDW	VDLM	VDF
bienfait		✗	✗
bonheur	✓	✓	✓
colère	✓	✓	✓
enthousiasme	✓	✓	✗
envie	✓	✓	✓
joie	✓	✓	✓
passion	✓	✓	✓
peur	✓	✓	✓
plaisir	✓	✓	✓
rire		✗	✗
sensation	✓	✓	✓
souvenir	✗	✓	✓
tendresse		✓	✓
tristesse	✓	✗	✓
rappel	0,91	0,79	0,79

TAB. 7.4 – Hyponymes du nom *émotion* dans les trois bases de voisins.

		VDW	VDLM	VDW
nb. de couples	sans seuil	7640	7590	5510
	seuil à 0,1	4960	4625	3827

TAB. 7.5 – Effets du seuil sur le nombre de couples d'hyponymes.

un hyperonyme et ses hyponymes. La mise au jour de ces contraintes nous permettra d'avoir une meilleure idée du type d'hyponymie que l'ADA est en mesure de repérer.

La démarche générale que nous adoptons ici consiste à comparer les hyperonymes de JDM en fonction de la proportion de leurs hyponymes qui sont captés par l'ADA. Pour cela, nous avons sélectionné un échantillon de couples en respectant la méthodologie que nous décrivons dans les 5 sous-sections suivantes.

7.3.1 Extraction des couples d'hyponymes

En suivant le protocole décrit à la section 7.2.2, nous avons successivement croisé les trois bases de voisins avec les hyponymes de JDM. Le nombre de couples contenu dans chacune de ces bases est rapporté au tableau 7.5 (ligne `sans seuil`).

7.3.2 Filtrage sur la productivité

Nous avons évoqué dans le préambule méthodologique (page 119) la tendance des mesures de similarité à rapprocher les mots qui ont des productivités comparables ainsi que les répercussions que cela peut avoir sur l'analyse des résultats. Ici encore, nous cherchons donc à écarter de notre échantillon les couples d'hypo/hyperonymes dont le rapport de productivté – Rprod – est particulièrement inégal. De la même façon que dans le chapitre 5, nous avons défini la valeur du Rprod pour les couples d'hypo/hyperonymes en nous appuyant sur la comparaison de cette valeur pour les couples de voisins et de non-voisins.

Nous avons représenté à la figure 7.1 l'évolution du nombre de couples d'hypo/hyperonymes voisins et non voisins en fonction de leur Rprod. Comme on peut le voir, les couples qui ont un Rprod inférieur à 0,17 ont peu de chances d'être captés par l'ADA (à la section 5.3.1, ce rapport était de 0,23 sur les données du DES). Dans la suite de cette étude, nous cherchons à étudier les phénomènes de décalage qui peuvent être imputés à des fonctionnements linguistiques. Or, comme on peut le voir ici, un grand nombre de

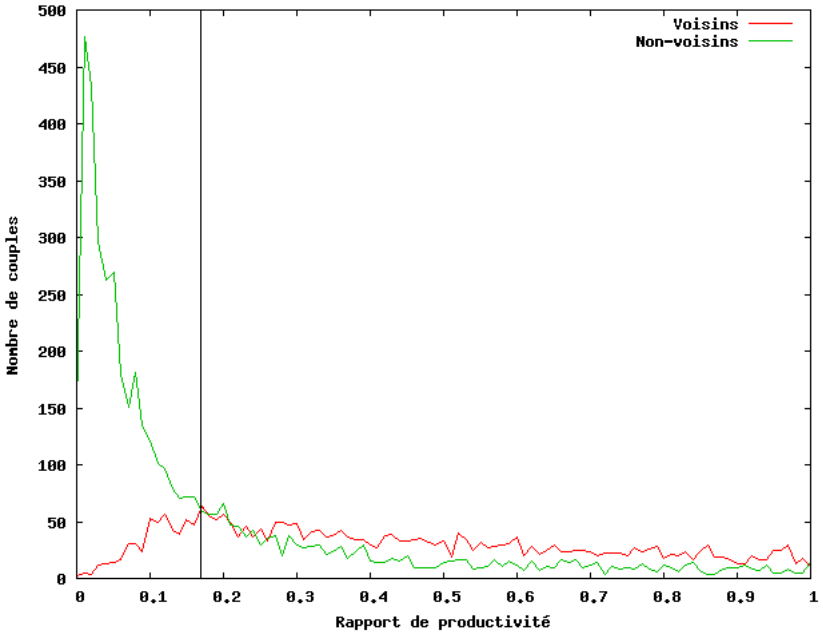


FIG. 7.1 – Évolution du nombre de couples voisins et non voisins en fonction du rapport de leurs productivités.

couples sont d'emblée défavorisés par l'ADA du fait que leurs productivités sont trop inégales. C'est la raison pour laquelle nous ne travaillons par la suite qu'avec des couples dont le rapport de productivité est supérieur ou égal à 0,1. Ce seuil nous semble à la fois assez restrictif – il met à l'écart la majorité des couples pour lesquels l'écart entre les productivités est le plus dommageable – et suffisamment permissif pour que nous puissions disposer d'un nombre raisonnable d'exemples à observer. Comme on peut le voir au tableau 7.5, cette manipulation fait considérablement baisser le nombre de couples d'hyponymes : la baisse est en moyenne de 35 %.

7.3.3 Mesure du rappel

Contrairement à l'étude que nous avons menée sur l'antonymie dans le chapitre précédent, nous avons ici fait le choix de ne pas étudier les couples d'hypo/hyperonymes indépendamment les uns des autres mais de les regrouper en fonction de leurs hyperonymes. En effet, comme nous l'avons vu à la section 6.1.2, l'antonymie est décrite comme une relation *one-to-one*, c'est-à-dire qu'un mot va s'opposer de façon beaucoup plus saillante avec un de ses antonymes potentiels qu'avec les autres. La relation d'hyperonymie ne

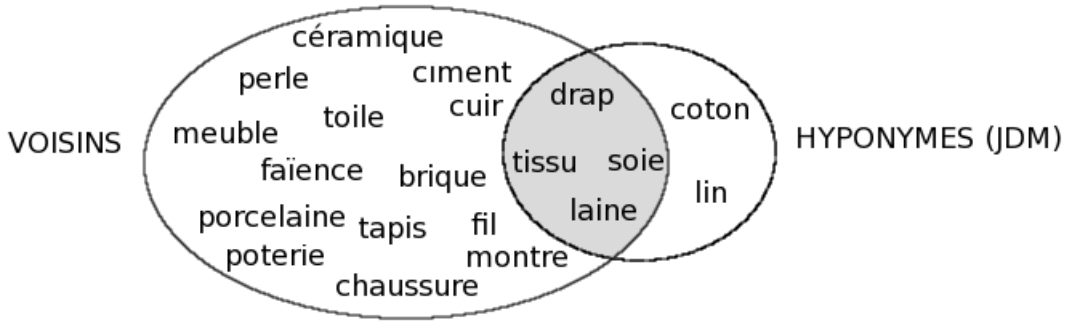


FIG. 7.2 – Illustration des notions de rappel et de précision à l'aide des voisins et des hyponymes du mot *étoffe*.

semble pas manifester ce type de fonctionnement de façon aussi marquée (la littérature ne mentionne pas d'hyponymie ou d'hyperonymie canonique, comme c'est le cas pour l'antonymie). C'est une relation *one-to-many* (Murphy, 2003) que nous avons choisi d'aborder comme telle en regroupant les couples d'hypo/hyperonymes par hyperonymes. Nous cherchons ainsi à voir s'il est possible de faire émerger des propriétés des hyperonymes en fonction de la proportion de leurs hyponymes qui sont captés par l'ADA.

Après avoir appliqué à nos couples d'hyponymes le seuil de productivité décrit précédemment, nous calculons donc, pour chaque hyperonyme, la proportion de ses hyponymes qui a été extraite par l'ADA (pour chacune des trois bases de voisins). C'est ce que nous appelons, à partir de maintenant, le *rappel*. Cette mesure, bien connue dans le domaine de la recherche d'information, va souvent de pair avec la mesure de précision. Nous nous appuyons sur la figure 7.2, qui rapporte les voisins et les hyponymes du mot *étoffe*, pour illustrer la façon dont nous utilisons ces notions dans la suite de ce travail :

- le rappel correspond ici au rapport entre les 4 hyponymes qui ont été captés par les voisins et le nombre total d'hyponymes de *étoffe* recensés dans JDM (soit 6), donc environ 0,67 ;
- la précision est le rapport entre les 4 voisins de *étoffe* qui sont recensés comme des hyponymes dans JDM et le nombre total de ses voisins (soit 18), autrement dit 0,22.

Dans le cadre de cette étude, nous nous focalisons sur le versant *rappel*. Nous privilégions donc par la suite des représentations comme celles qui figurent aux tableaux 7.6, dans lesquels n'apparaissent pas le nombre total de voisins de l'hyperonyme observé. On peut y voir que parmi les 9 hyponymes de *machine*, 7 ont été captés par l'ADA (voisins de Wikipédia). Cela correspond à un rappel d'environ 0,78. Dans le cas du nom *artisan*, le tableau montre

machine	
hyponyme	voisins ?
four	✗
lanceur	✗
engin	✓
hélicoptère	✓
locomotive	✓
métier	✓
moteur	✓
moulin	✓
ordinateur	✓
rappel	0,78

artisan	
hyponyme	voisins ?
facteur	✗
maçon	✗
mécanicien	✗
menuisier	✗
voilier	✗
ouvrier	✓
peintre	✓
rappel	0,29

TAB. 7.6 – Variation du repérage des hyponymes des noms *machine* et *artisan* dans les VDW.

	avant filtrage			après filtrage		
	VDW	VDLM	VDF	VDW	VDLM	VDF
nb. d'hypero.	1246	1278	888	346	311	210
nb. moyen d'hypo. par hypero.	4	4	4	11	10	12
rappel moyen	0,58	0,55	0,34	0,60	0,58	0,34

TAB. 7.7 – Quelques propriétés des hyperonymes de JDM.

que seulement 2 des 7 hyponymes ont été extraits. Le rappel est d'environ 0,29.

7.3.4 Filtrage en fonction du nombre de voisins et d'hyponymes

On peut voir, dans la partie avant filtrage du tableau 7.7, que le nombre d'hyperonymes présents dans les 3 bases de voisins varie de 888 – dans les VDF – à 1278 – dans les VDLM. Cela est dû au fait que les corpus à partir desquels ont été calculées les bases de voisins ont des lexiques de taille et de nature différentes. Le rappel lui aussi varie entre les bases : c'est encore une fois dans les VDF qu'il est le plus faible.

En revanche, le nombre d'hyponymes moyen par hyperonyme reste stable. Le fait que la moyenne soit de 4 peut laisser supposer qu'un grand nombre d'hyperonymes n'ont qu'un ou deux hyponymes. C'est en effet le cas pour environ 66 % d'entre eux. Ce constat est problématique pour nous dans la

mesure où le phénomène que nous cherchons à analyser s'observe d'autant mieux que le nombre d'hyponymes d'un mot est élevé. C'est la raison pour laquelle nous avons fait le choix de ne conserver que les hyperonymes qui ont au moins 4 hyponymes. Nous avons également mis un seuil de 4 sur le nombre de voisins par hyperonyme puisque si un hyperonyme n'a qu'un ou deux voisins, son rappel sera nécessairement faible (ce seuil n'a cependant qu'une très faible influence sur le nombre d'hyperonymes).

On peut voir, dans le tableau 7.7, que ce filtrage réduit en moyenne de 75 % le nombre d'hyperonymes de chaque ressource. Comme on pouvait s'y attendre, le nombre moyen d'hyperonymes par mot est quasiment multiplié par 3. Le rappel, en revanche, varie assez peu.

7.3.5 Gérer la variation entre les ressources

À ce stade, nous avons appliqué aux données de notre échantillon les filtres suivants :

1. nous avons imposé aux couples un Rprod minimum de 0,1 ;
2. après avoir calculé le rappel de chaque hyperonyme, nous avons choisi de ne conserver que ceux qui ont au moins 4 voisins et 4 hyponymes.

Nous présentons maintenant les mesures que nous avons utilisées pour caractériser la façon dont sont repérés les hyponymes d'un hyperonyme dans les différents corpus.

Nous avons rapporté au tableau 7.8, pour chacune des trois ressources étudiées, 10 des hyperonymes pour lesquels aucun hyponyme n'a été capté par les voisins (leur rappel est donc de 0). Étant donné que le phénomène de décalage entre la ressource et les voisins est d'autant plus visible que l'hyperonyme a plus d'hyponymes, nous avons sélectionné ici les hyperonymes qui ont le nombre le plus élevé d'hyponymes. Ainsi, dans notre tableau, le mot *matière* illustre le cas de décalage le plus extrême, dans le sens où 22 de ses hyponymes sont présents dans le corpus Frantext mais aucun d'entre eux n'a été capté comme un de ses voisins distributionnels.

Nous avons montré à la section 7.2.3 que la composition des corpus à partir desquels ont été calculées les bases de voisins pouvait avoir une influence sur le type et la proportion d'hyponymes captés. Dans l'échantillon présenté au tableau 7.8, on peut voir que deux hyperonymes – *pomme* et *orientation* – ont un rappel nul à la fois dans les VDW et les VDLM. Le fait que le décalage distributionnel entre ces deux mots et leurs hyponymes s'observe aussi bien pour les VDW que pour les VDLM peut nous laisser penser que l'on a affaire a un phénomène qui n'est pas propre à un type de corpus particulier. On peut alors supposer que l'on a affaire à un phénomène plus généralisable, qui

VDW		VDLM		VDF	
hypero	nb hypo	hypero	nb hypo	hypero	nb hypo
préfecture	17	*accessoire*	13	*matière*	22
pomme	9	*vivant*	11	*prénom*	22
orientation	8	*orientation*	8	*activité*	16
verbe	6	*relief*	8	*viande*	14
épice	6	*extérieur*	6	*mesure*	13
périphérique	6	*pomme*	6	*individu*	11
teinture	6	*textile*	6	*jeu*	9
échecs	5	*accompagnement*	6	*roche*	8
embarcation	5	*issue*	6	*ton*	8
anatomie	5	*gâteau*	6	*direction*	8

TAB. 7.8 – Exemples de mots ayant des rappels nuls dans les trois bases étudiées.

nous permet de mettre en lumière des types de couples d'hypo/hyperonymes pour lesquels le critère de substituabilité ne s'applique tout simplement pas.

De ce fait, il est important, pour la suite de notre étude, de gérer ces effets liés à la nature du corpus afin de pouvoir décrire de façon distincte les cas de décalage qui s'observent :

- de façon stable, quelle que soit la base de voisins utilisée ;
- de façon fluctuante d'une base de voisins à l'autre.

Pour cela, nous avons extrait, à partir des jeux de couples d'hypo/hyperonymes précédemmment constitués, les 162 hyperonymes qui apparaissent dans chacune des trois bases de voisins. Pour chacun d'entre eux, nous avons calculé :

- le rappel moyen ;
- l'écart type au rappel.

Le rappel moyen nous donne une indication des proportions dans lesquelles les hyponymes d'un mot ont été captés dans les trois bases. L'écart type nous indique dans quelle mesure le rappel varie d'une ressource à l'autre. L'écart type correspond à la moyenne des écarts, en valeurs absolues, entre le rappel calculé pour chaque corpus et le rappel moyen. Autrement dit, soient r_{vd*} le rappel calculé pour $_{hypero}$ dans les voisins de * et $\bar{r}$ le rappel moyen, l'écart type σ_{hypero} se calcule de la façon suivante :

$$\sigma_{hypero} = \frac{|r_{vdw} - \bar{r}| + |r_{vdlm} - \bar{r}| + |r_{vdf} - \bar{r}|}{3}$$

Plus l'écart type est haut, plus le rappel varie entre les ressources. Ainsi, pour reprendre les exemples des tableaux 7.3 et 7.4, *engin* a un rappel moyen de 0,40 et un écart type de 0,36 alors que *émotion* a un rappel moyen de 0,83 et un écart type de 0,06. Cela signifie que le repérage des hyponymes se fait dans des proportions plus équivalentes dans les trois corpus pour *émotion* que pour *engin*.

Afin de voir si le rappel des hyperonymes tend vers la stabilité ou s'il varie fortement d'une ressource à l'autre, nous avons calculé la moyenne de leurs écarts types. Ce dernier est de 0,2. Cela signifie que le rappel d'un hyperonyme varie en moyenne d'environ 0,2 point entre les trois bases de voisins. Ce degré de variation est relativement faible. Il est révélateur d'une certaine stabilité dans le repérage des hyponymes entre les corpus.

7.4 Analyse du décalage distributionnel entre les hyperonymes et leurs hyponymes

Dans la section précédente, nous avons appliqué une série de filtres aux couples d'hypo/hyperonymes de JDM afin d'écarter ceux qui sont par avance défavorisés par le calcul distributionnel (cf. préambule méthodologique). Nous avons ensuite regroupé les couples comportant le même hyperonyme. Notre étude se voulant comparative, nous avons choisi de n'étudier que les 162 hyperonymes qui apparaissent dans les trois bases de voisins. Dans cette section, nous nous livrons à l'analyse de ces données.

Nous avons vu que ces hyperonymes variaient selon deux critères, à savoir le rappel moyen et l'écart type. Dans les sections qui suivent, nous étudions successivement trois ensembles d'hyperonymes que nous avons dégagés en fonction de ces deux valeurs, comme le montre le tableau ci-dessous.

<table>
<tr><th></th><th>écart type élevé</th><th>écart type faible</th></tr>
<tr><td>rappel élevé</td><td rowspan="2">section 7.4.3</td><td>section 7.4.1</td></tr>
<tr><td>rappel faible</td><td>section 7.4.2</td></tr>
</table>

Nous commençons donc par étudier les hyperonymes pour lesquels le rappel est élevé et l'écart type faible (section 7.4.1). Ces hyperonymes sont ceux dont les hyponymes ont été captés en masse, et ce dans les trois bases de voisins. Le deuxième cas est celui sur lequel nous nous attardons le plus. Il s'agit des hyperonymes qui ont un rappel et un écart type faibles, ou, autrement dit, ceux dont les hyponymes sont mal repérés dans les trois bases de voisins (section 7.4.2). Le troisième et dernier cas de figure est celui des

hyperonymes dont les hyponymes ont été extraits de façon variable en fonction des bases de voisins (section 7.4.3). Bien que de tailles modestes – 26 hyperonymes en moyenne –, ces trois ensembles nous donnent à voir trois modes de fonctionnement différents du rapport de substituabilité entre les hyponymes et leurs hyperonymes.

7.4.1 Rappel élevé dans les trois ressources

Dans un premier temps, nous nous intéressons aux hyperonymes dont les hyponymes ont été bien repérés, et ce dans les trois bases de voisins. Nous avons choisi d'extraire les 30 hyperonymes dont le rappel est le plus élevé et dont l'écart type est supérieur ou égal à 0,21 (soit la moyenne des écarts types). Ces hyperonymes ont été rapportés au tableau 7.9.

Les hyperonymes dont les hyponymes ont été les mieux captés sont *parent* et *peur*. Nous abordons ces deux cas en particulier par la suite.

Parmi les hyponymes de *parent*, on peut citer *neveu*, *oncle*, *cousin*, *mère*, *frère*, *père* ou *fils*. Parmi les contextes qui ont permis de rapprocher *parent* de ses hyponymes – ils sont communs à au moins 6 des 7 hyponymes que nous venons de citer –, on trouve principalement la position sujet de verbes comme *mourir*, *fonder*, *envoyer*, *décider* ou *occuper* et des prédicats nominaux comme *mariage*_DE, *histoire*_DE, *ami*_DE, *mère*_DE ou *maison*_DE. Les adjectifs, en revanche, semblent n'avoir joué qu'un rôle mineur dans le rapprochement de *parent* et de ses hyponymes.

On note que, à l'instar de *parent*, 11 des 30 hyperonymes listés dans le tableau 7.9 renvoient à des noms d'animés humains : *parent*, *enfant*, *soldat*, *homme*, *officier*, *artiste*, *chef*, *employé*, *femme*, *prince*, *personne*. Toutefois, les modalités qui ont fait que ces hyperonymes aient été rapprochés de leurs hyponymes ne sont pas forcément les mêmes que celles qu'on a pu observer pour *parent*. En effet, alors que les contextes adjectivaux n'intervenaient que très peu dans le rapprochement de *parent* et de ses hyponymes, on constate qu'ils jouent un rôle primordial dans le cas d'un hyperonyme comme *artiste*. Les contextes qui permettent de rapprocher ce mot avec ses hyponymes *compositeur*, *musicien*, *poète*, *peintre*, etc. sont pour la plupart des adjectifs de nationalité (*québécois*, *américain*, *irlandais*) ou autres (*né*, *reconnu*, *originaire* ou *meilleur*). Dans le cas de l'hyperonyme *soldat*, il est difficile de dégager une tendance dans les contextes qui ont permis de rapprocher ce mot à ces hyponymes (*colonel*, *lieutenant*, *chevalier*, *capitaine*, etc.). On trouve, dans des proportions sensiblement équivalentes, des contextes adjectivaux (*premier*, *allemand*, *jeune*), verbaux – principalement sujet (*mener*, *demander*, *mourir*) – ou nominaux (*grade*_DE, *qualité*_DE, *rôle*_DE). Autrement dit, bien qu'on puisse clairement constater une tendance des hyperonymes

hyperonyme	rappel	écart type
parent	1	0
peur	1	0
enfant	0,95	0,07
soldat	0,93	0,1
type	0,93	0,09
homme	0,93	0,05
petit	0,92	0,12
livre	0,92	0,12
officier	0,92	0,12
artiste	0,91	0,07
chef	0,89	0,16
employé	0,88	0,09
trace	0,87	0,19
ouvrage	0,87	0,19
aventure	0,87	0,19
femme	0,87	0,02
émotion	0,85	0,05
salle	0,83	0,13
sentiment	0,8	0,06
prince	0,77	0,21
habit	0,77	0,15
son	0,77	0,07
métal	0,75	0,18
pensée	0,74	0,07
corps	0,74	0,05
école	0,72	0,16
temps	0,72	0,06
vue	0,69	0,2
visage	0,68	0,14
personne	0,68	0,05

TAB. 7.9 – Les 30 hyperonymes dont les hyponymes sont les mieux captés dans les trois bases de voisins.

renvoyant à des noms d'humains à partager les contextes d'apparition de leurs hyponymes, il est difficile de dégager des fonctionnements qui semblent s'appliquer à l'ensemble des exemples des 11 exemples de notre échantillon.

L'hyperonyme *peur*, dont les hyponymes ont également été tous captés – dans les trois bases de voisins –, est plus litigieux. Le nom *peur* n'a que 4 hyponymes, ce qui correspond au seuil que nous avions fixé (cf. section 7.3.5). Ces derniers sont *terreur*, *horreur*, *crainte* et *angoisse*. Nous considérons ce cas comme litigieux dans le sens où les couples *peur*/*terreur*, *peur*/*horreur*, *peur*/*crainte* et *peur*/*angoisse* semblent davantage relever de la synonymie, et sont d'ailleurs considérés comme des synonymes dans le DES.

On remarque que parmi les 30 hyperonymes du tableau 7.9, *sentiment* et *émotion* sont sémantiquement proches de *peur*. Le caractère catégorisant des noms *sentiment* et *émotion* apparaît toutefois plus intuitif que pour *peur*. On note que ces deux mots partagent un grand nombre de leurs hyponymes : *bonheur*, *joie*, *passion*, *peur*, *plaisir*, *rire*, *souvenir*, *tendresse*, *tristesse*. Cela n'est pas surprenant outre mesure étant donné que ces deux mots sont clairement synonymes. On peut donc s'attendre à ce qu'ils aient des fonctionnement analogues dans le corpus et, de ce fait, il paraît normal de les retrouver tous deux dans notre échantillon.

Il est difficile de caractériser le reste des hyperonymes de notre échantillon. Ces derniers renvoient en effet à des concepts de natures hétérogènes, aussi bien abstraits (*temps*, *vue*, *aventure*) que concrets (*ouvrage*, *habit*, *visage*). Ici encore, on peut être surpris par le fait que certains de ces mots apparaissent comme des catégorisants. On a ici affaire à des conceptions de la relation d'hyperonymie qui sont :

- particulièrement vagues, comme pour l'hyperonyme *temps*, dont les hyponymes sont soit des unités de temps (*heure*, *mois*, *jour*, *siècle*, *année*), soit des mots qui renvoient aux champ lexical du climat (*chaleur*, *pluie*, *soleil*) ;
- erronées, comme dans le cas de *visage*, dont les hyponymes sont des synonymes (*face*, *figure*) ou des méronymes (*nez*, *bouche*, *œil*).

Ces deux cas de figure illustrent les limites de l'utilisation d'une ressource comme JDM, que nous avons évoquées dans le préambule méthodologique.

Les résultats de l'étude de ce premier échantillon nous montrent donc une tendance manifeste des hyperonymes renvoyant à des animés humains à partager les contextes de leurs hyponymes. Ces premières observations ne nous ont toutefois pas permis de mettre au jour une régularité dans la nature des contextes partagés (on aurait par exemple pu penser que la position sujet des verbes favoriserait les rapprochements de noms d'humains). On peut toutefois supposer que le fait que ce phénomène s'observe dans les trois bases distributionnelles soit lié à la nature des corpus Wikipédia, Le Monde

et Frantext, dans lesquels les individus tiennent une place importante.

7.4.2 Rappel faible dans les trois ressources

Dans cette section, nous étudions les hyperonymes qui sont les plus représentatifs du phénomène de décalage *en langue*, que nous avons définis comme étant ceux pour lesquels les hyponymes sont mal captés dans les trois bases de voisins. Nous avons donc extrait de notre jeu de 162 hyperonymes ceux dont :

- au maximum un tiers des hyponymes a été capté (le rappel est inférieur ou égal à 0,33). Ce seuil correspond à un compromis entre l'impératif d'étudier des cas où le décalage est le plus marqué et la nécessité de disposer de suffisamment d'exemples à analyser (plus le seuil est bas, plus le nombre d'hyperonymes est faible) ;
- l'écart type est inférieur ou égal à 0,21 (soit la moyenne des écarts types).

Ce double filtrage nous permet de recueillir l'échantillon de 19 hyperonymes rapporté au tableau 7.10. La taille modeste de cet échantillon est due aux multiples filtrages que nous avons dû appliquer aux données de départ. Ces derniers se sont toutefois avérés nécessaires pour l'extraction des hyperonymes les plus représentatifs du phénomène que l'on souhaite étudier.

Nous décrivons maintenant les causes de décalage entre hyperonymes et hyponymes que l'analyse de ces exemples nous a permis de mettre au jour.

7.4.2.1 Les hyponymes sont polysémiques

On a ici affaire à des mots dont les hyponymes renvoient principalement à des mots polysémiques/homonymiques :

- *poisson* : *empereur, lieu, perche, bar, clown*, etc.
- *pomme* : *jazz, gala, canada, empire, tentation*, etc.
- *oiseau* : *fou, pape, veuve, pic, secrétaire*, etc.

Dans le cas de *poisson*, les sens de *empereur*, *bar*, et *clown* en tant qu'hyponymes de *poisson* n'émergent pas dans leurs distributions. C'est-à-dire que, contrairement à *saumon*, qui est capté comme un voisin de *poisson*, ils n'apparaissent pas dans des contextes qui lui sont caractéristiques, comme les modifieurs *mariné*, *frais*, *sauvage* ou *fumé* (VDLM).

Le problème que l'on rencontre – et que l'on a déjà eu l'occasion d'évoquer (cf. section 2.1.1) – est celui de la sparsité (*sparseness*). Il se manifeste ici par le fait que, pour un mot polysémique, l'une de ses acceptions est tellement rare dans le corpus qu'elle n'émerge pas dans la distribution du mot. Par

hyperonyme	rappel	écart type
pomme	0	0
prénom	0,06	0,05
viande	0,12	0,1
parler	0,13	0,09
vivant	0,14	0,12
meuble	0,17	0,13
grade	0,17	0,16
plante	0,18	0,11
médaille	0,18	0,14
oiseau	0,21	0,09
poisson	0,22	0,05
sens	0,23	0,09
repos	0,25	0,2
roche	0,26	0,2
métier	0,27	0,16
unité	0,28	0,21
toit	0,29	0,03
ballon	0,32	0,17
substance	0,32	0,2

TAB. 7.10 – Les 19 hyperonymes dont les hyponymes sont les moins captés dans les trois bases de voisins.

exemple, dans le cas de *perche*, les emplois qui prévalent dans les corpus renvoient aux deux acceptions suivantes (ici aussi extraites du TLF) :

- "Longue pièce de bois ou de métal mince, à peu près ronde et de faible section"
  Exemples de contextes : <*perche*,MOD,*vertical*>, <*tendre*,OBJ,*perche*>, <*chalutier*,À,*perche*>[3] ;
- "Tige flexible [...] utilisée par les athlètes qui pratiquent un exercice de saut en hauteur"
  Exemples de contextes : <*saut*,À,*perche*>, <*battre*,À,*perche*>, <*perche*, EN,*salle*>, <*perche*,MOD,*féminin*>. Il est à noter que la plupart de ces contextes renvoient en fait au syntagme *saut à la perche*, qui est quelquefois remplacé par *perche* seul, comme dans le contexte *battre à la perche*.

Il est possible de trouver quelques contextes – principalement dans le corpus Wikipédia – où *perche* est employé en tant qu'hyponyme de *poisson* : <*perche*,DE,*amérique*>, <*perche*,DE,*mer*>, <*perche*,DE,*nil*>. Ces derniers restent cependant trop marginaux pour que cette acception puisse émerger (et qu'on puisse observer un recouvrement de ses contextes avec ceux de *poisson*).

Le problème que nous rencontrons ici a notamment été évoqué dans Véronis (2003, p. 2) :

> Les techniques basées sur les vecteurs de mots se heurtent toutefois à une difficulté majeure et rédhibitoire : la très grande différence de fréquence entre usages d'un même mot (déjà constatée par Zipf, 1945) repousse la plupart des distinctions utiles endessous du seuil de bruit du modèle. Ainsi, selon nos estimations, l'usage "match de barrage" concerne moins de 1 % des documents contenant le mot *barrage*.

7.4.2.2 Les hyperonymes renvoient à des facettes sous-représentées

Le cas de figure que nous présentons ici ressemble assez au cas précédent. Nous expliquons en effet le décalage distributionnel entre des hyperonymes comme *viande*, *plante* ou *substance* et leurs hyponymes par le fait qu'ils réfèrent à des sens qui ne se manifestent que de façon marginale dans le corpus. La différence est qu'ici, comme nous le montrons plus loin, nous n'avons pas vraiment affaire à des entités polysémiques.

[3]Malgré la proximité sémantique de *chalutier* et de *perche* – en tant qu'hyponyme de *poisson* – c'est bien le sens de "Longue pièce de bois ou de métal mince" qui est actualisé ici.

viande			
	VDW	VDLM	VDF
agneau	✗		
araignée	✗		✗
blanc	✗	✗	✗
bœuf		✓	✗
canard	✗	✗	✗
cochon	✗		✗
culotte			✗
épaule	✗	✗	✗
filet	✗	✗	✗
foie	✓	✓	
lapin	✗	✗	✗
mouton	✗	✗	✗
oie	✗		
porc	✗	✓	✗
poulet	✓	✗	✗
sanglier	✗		
souris	✗	✗	✗
veau	✗	✗	✗
rappel	0,12	0,25	0

TAB. 7.11 – Hyponymes du nom *viande* dans les voisins de Wikipédia (VDW), Le Monde (VDLM) et Frantext (VDF).

viande				
proportion (%)	contexte			compatible avec l'hyperonyme ?
	mot	cat	rel.	
75	*petit*	A	_	✗
50	*noir*	A	_	✗
50	*voir*	V	OBJ	✗
44	*foie*	N	DE	✗
44	*élevage*	N	DE	✗
44	*sauvage*	A	_	✗
38	*peau*	N	DE	✗
38	*élever*	V	OBJ	✗
38	*tête*	N	DE	✗
31	*viande*	N	DE	✗

TAB. 7.12 – Les 10 contextes partagés par le plus grand nombre d'hyponymes de *viande* dans les VDW. Les chiffres de la première colonne représentent la proportion d'hyponymes qui partagent ce contexte.

Nous avons rapporté au tableau 7.11 l'ensemble des hyponymes de *viande* dans les trois ressources[4]. On peut voir que la plupart d'entre eux renvoient à des noms d'animaux. Ces derniers ont été catégorisés comme des hyponymes de *viande* étant donné qu'ils peuvent être employés dans le sens "chair comestible d'un animal". Comme nous allons le montrer ici, le décalage distributionnel entre *viande* et ses hyponymes est dû au fait que ce sens n'est pas représenté dans les corpus.

Le tableau 7.12 rapporte les contextes d'apparition les plus partagés par les hyponymes de *viande* dans les VDW[5]. La première colonne indique le pourcentage d'hyponymes de *viande* qui apparaissent dans les contextes donnés. Les croix rouges dans la dernière colonne indiquent qu'aucun d'entre eux n'est partagé par *viande*. La raison en est que ces contextes réfèrent à *canard*, *mouton*, etc. en tant qu'animaux (et la viande n'est pas un animal). Par exemple, on voit que 75 % des hyponymes sont modifiés par l'adjectif *petit* mais que ce n'est pas le cas de *viande*, leur hyperonyme.

Toutefois, on peut voir au tableau 7.11 que *poulet*, *porc* et *bœuf* partagent suffisamment de leurs contextes avec *viande* pour être captés par l'ADA. Ces noms d'animaux sont en effet les plus susceptibles d'être employés comme des

[4]Comme nous l'avons vu à la section 7.2.3, le fait que de nombreux mots soient absents des trois bases s'explique par le fait qu'elles ont subi plusieurs seuillages.

[5]Pour des raisons de commodité, nous ne fournissons ici que des exemples issus d'une seule base. Le phénomène reste cependant le même dans les deux autres.

viande				
proportion (%)	contexte			compatible avec l'hyperonyme ?
	mot	cat	rel.	
31	*manger*	V	OBJ	✓
31	*utiliser*	V	OBJ	✓
25	*trouver*	V	OBJ	✓
19	*gras*	A	_	✓
19	*morceau*	N	DE	✓
12	*ajouter*	V	OBJ	✓
12	*donner*	V	OBJ	✓
12	*ragoût*	N	DE	✓
12	*agneau*	N	_	✓
12	*poulet*	N	_	✓

TAB. 7.13 – Les 10 contextes de *viande* les mieux partagés par ses hyponymes.

entités consommables dans nos corpus. Nous avons rapporté au tableau 7.13 les contextes compatibles avec *viande* les plus partagés avec ses hyponymes.

Il est intéressant de noter que le mot *viande* apparaît parmi les contextes les mieux partagés par ses hyponymes (contexte *viande*_DE dans le tableau 7.12). Ce dernier sert à activer un sens qui n'est donc pas présent par défaut dans les occurrences de *mouton*, *lapin*, etc. dans le corpus Wikipédia. Il traduit de par là même un fonctionnement syntagmatique de la relation d'hyperonymie.

Contrairement aux cas comme *poisson*, que nous avons évoqués plus tôt, l'exemple que nous venons de développer peut être distingué de la polysémie à proprement parler. En effet, les sens "animal" et "chair comestible d'un animal" de mots comme *poulet* ou *veau* correspondent à ce que Cruse (2002, 2004, 2006) ou encore Croft et Cruse (2004) appellent des *facettes*. La différence se situe au niveau du fait que les différentes facettes d'un mot peuvent être compatibles entre elles alors que ce n'est pas le cas de deux acceptions d'un mot polysémique :

(4) Il a tué son lapin et l'a cuisiné en civet.

(5) ? Il a pêché un bar et y a bu une bière.

On peut en effet voir que la cohabitation des deux sens de *bar* dans l'exemple (5) produit un zeugma alors que ce n'est pas le cas pour les deux facettes de *lapin* dans l'exemple (4).

Nous expliquons ainsi le décalage entre *plante* et ses hyponymes *salade*,

riz, *tabac*, *rose* etc. Par exemple, on peut voir que, dans le corpus, l'analyse des contextes dans lesquels apparaissent *salade* et *tabac* font émerger des facettes qui sont sous-représentées dans la distribution de *plante* :

- le premier voisin de *salade* dans les VDW est *plat*, avec lequel il partage la position objet de verbes comme *parfumer*, *aromatiser*, *servir* ou *manger*. Autrement dit, dans le corpus Wikipédia, le nom *salade* est majoritairement employé pour désigner quelque chose que l'on peut consommer ;
- le nom *tabac* n'apparaît pas non plus dans les mêmes contextes que *plante*. Il partage avec *coton*, *chaussure* et *drap* des contextes comme *manufacture*_DE, *production*_DE ou *entrepot*_DE qui montrent que le tabac est considéré dans le corpus comme un produit industriel.

Bien que, *stricto sensu*, les salades et le tabac soient des plantes, on voit que, dans ce cas, les emplois qui sont faits dans le corpus entre l'hyperonyme et ses hyponymes divergent : *salade* et *tabac* ne sont que minoritairement employés pour désigner des végétaux. Ainsi, on voit que la façon dont le corpus *modèle* la représentation des mots que l'ADA permet de faire émerger peut entrer en conflit avec une représentation de type taxonomique établie *in abstracto*.

7.4.2.3 Usage *vs* mention

Premier cas de figure : *prénom* On peut voir au tableau 7.10 que le rappel de *prénom* est pratiquement nul. Et si l'on s'intéresse aux trois prénoms qui ont été captés comme des voisins de *prénom*, on s'aperçoit qu'on a affaire à du bruit :

- *roman* partage avec *prénom* des contextes comme se *terminer*_SUJ, *traduction*_DE ou *écrire*_OBJ ;
- *marine* et *victoire*, comme *prénom*, sont modifiés par un ensemble d'adjectifs de nationalité (*japonais, portugais, espagnol*, etc.).

La raison du décalage entre la distribution de *prénom* et celle de ses hyperonymes – *benjamin*, *charles*, *ferdinand*, etc. – est à attribuer à la nature même des objets auxquels ils réfèrent. En effet, le mot *prénom* a une valeur métalinguistique alors que ses hyponymes désignent des êtres animés. Les exemples suivants illustrent ces deux types d'emplois :

(6) J'aime bien Sankar. **Ce prénom** me rappelle ma jeunesse à Pondichéry.

(7) J'aime bien Sankar. **Ce garçon** me rappelle mon cousin de Pondichéry.

prénom				
proportion (%)	contexte			compatible avec l'hyperonyme ?
	mot	cat	rel.	
60	*saint*	A	_	✗
60	*jeune*	A	_	✗
48	*bourbon*	N	_	✗
48	*petit*	A	_	✗
44	*II*	N	_	✗
36	*voir*	V	OBJ	✓
36	*I*	N	_	✗
36	*premier*	A	_	✓
32	*france*	NP	_	✗
32	*grand*	A	_	✗

TAB. 7.14 – Les 10 contextes partagés par le plus grand nombre d'hyponymes de *prénom* dans les VDW.

Dans l'exemple (6), *Sankar* est employé *en mention*, il fait référence au signe linguistique (on aurait tendance à l'italiciser). L'exemple (7), en revanche, illustre un emploi *en usage*, c'est-à-dire que *Sankar* réfère à un individu. Les contextes d'apparition des hyponymes de *prénom* nous indiquent qu'ils sont massivement employés en usage dans nos corpus. En effet, ils ne partagent pas la plupart des contextes de *prénom* comme *diminutif*_DE, *syllabe*_DE, *orthographier*_OBJ ou *étymologie*_DE, qui sont caractéristiques d'emplois métalinguistiques. Réciproquement, on peut voir au tableau 7.14 que les contextes les mieux partagés par les hyponymes de *prénom* sont, pour la plupart, incompatibles avec le mot *prénom*.

De par sa taille réduite, notre échantillon ne nous a permis de décrire qu'un seul exemple illustrant ce type de décalage métalinguistique. Toutefois, nous avons pu tirer les mêmes conclusions de notre observation de la distribution d'hyperonymes comme *lettre*, dont les hyponymes – *x*, *v*, *n*, etc. – ne partagent pas les contextes d'apparition, ou encore *consonne* et *verbe*. Cela nous incite à penser que ce phénomène constitue bel et bien une source de décalage pour les mots du métalangage qui désignent des unités linguistiques.

Deuxième cas de figure : *grade* et *métier* Les hyperonymes *grade* et *métier* ont la particularité de ne pas présenter les propriétés typiques de l'hyperonymie. Il leur est notamment impossible d'entrer dans les patrons *Un/le X est un (type de) Y* :

(8) Un lapin est un animal.

(9) * Un facteur est un métier.

(10) * Un colonel est un grade.

Pour que les exemples (9) et (10) soient grammaticaux, il faut modifier le patron de la façon suivante :

(11) *Facteur* est un métier.

(12) *Colonel* est un grade.

Autrement dit, il faut supprimer le déterminant et recourir à un emploi en mention de *facteur* et *colonel*. Ce fonctionnement rappelle celui de *prénom*, évoqué plus haut. La différence ici est que les hyponymes sont des noms communs. De plus, le critère de substituabilité ne semble pas s'appliquer pour ces deux mots :

(13) Ce laboratoire abrite des chercheurs.

n'implique pas

(14) * Ce laboratoire abrite des métiers.

De la même façon que :

(15) Le général est venu saluer les troupes.

n'implique pas

(16) * Le grade est venu saluer les troupes.

Pourtant, un locuteur reconnaîtra de façon instinctive *poissonnier*, *facteur* ou *secrétaire* comme des types de métier, ou encore *colonel*, *général* et *lieutenant* comme des types de grades. Nous tentons ici de voir ce que peuvent nous apprendre les voisins sur ce paradoxe apparent.

Nous avons rapporté au tableau 7.15 les contextes que partagent le plus grand nombre d'hyponymes de *grade*. Le seul qui soit compatible avec l'hyperonyme est *armée* (dans les cas où *grade* est prédicat et porte la relation DE). On note que 4 de ces 10 contextes sont des positions sujet de verbes d'action, dans lesquels *grade* n'apparaît pas. Cela nous indique que *sergent*, *major*, *capitaine*, etc. sont employés dans le corpus comme des entités animées. De la même façon, les contextes *ordre*_DE et *commandement*_DE appellent un nom qui renvoie à un animé. Ainsi, le décalage qui s'observe dans le cas de *grade* et *métier* peut s'expliquer par le fait que ces mots ont pour hypo-

grade				
proportion (%)	contexte			compatible avec l'hyperonyme ?
	mot	cat	rel.	
100	*commander*	V	SUJ	✗
100	*diriger*	V	SUJ	✗
100	*armée*	N	_	✓
100	*britannique*	A	_	✗
100	*faire*	V	SUJ	✗
88	*grade*	N	DE	✗
88	*garde*	N	_	✗
88	*commandement*	N	DE	✗
88	*ordre*	N	DE	✗
88	*recevoir*	V	SUJ	✗

TAB. 7.15 – Les 10 contextes partagés par le plus grand nombre d'hyponymes de *grade* dans les VDW.

nymes des mots qui sont de natures différentes de la leur. Leurs distributions divergent donc fortement, ce qui compromet le repérage de paires comme *grade/major*, *grade/sergent* ou *métier/arbitre*, *métier/juge* par l'ADA.

Afin d'illustrer ce décalage, nous avons rapporté au tableau 7.16 les 10 contextes dans lesquels apparaît le plus fréquemment le mot *grade* dans le corpus Wikipédia. Ces derniers sont majoritairement incompatibles avec ses hyponymes. Il est à noter toutefois qu'une proportion non négligeable d'hyponymes de *métier* a été captée dans les VDW et les VDLM (*compositeur*, *médecin*, *ingénieur*, etc.). On remarque cependant que, dans la plupart des cas, ces rapprochements sont dus à des contextes communs qui sont des modifieurs très productifs comme *premier*, *principal*, *nouveau*, *ancien*, *nombreux*, etc.

7.4.2.4 Relations non hyperonymiques

Nous concluons cette section en évoquant les quelques hyperonymes de notre échantillon qui présentent un cas de décalage qui peut être expliqué par le fait que leurs hyponymes ne constituent pas, pour la plupart, ce que l'on pourrait réellement considérer comme des hyponymes. Nous avons en effet montré, à la section 7.2.1 que les mécanismes de jeu de JDM induisaient des biais dans le type de relations qui sont produites. C'est notamment le cas des exemples suivants :

– *médaille* : *militaire, or, bronze, argent, récompense*, etc.

mot	cat.	relation	fréquence
promouvoir	V	à	365
obtenir	V	obj	311
élever	V	à	197
liste	N	de	127
recevoir	V	obj	103
conférer	V	obj	88
valoir	V	obj	86
insigne	N	de	83
atteindre	V	obj	81
monter	V	en	74

TAB. 7.16 – Les 10 contextes d'apparition les plus fréquents du mot *grade* dans les VDW.

– *ballon* : *plastique, cuir, rugby, sport, football*, etc.

On peut voir que dans les mots qui sont recensés dans JDM comme des hyponymes de *médaille* et *ballon* relèvent soit d'une relation thématique (*médaille/récompense*, *ballon/football*, etc.), soit de la méronymie (*médaille/or*, *ballon/plastique*, etc.). Nous imputons ces erreurs à une possible incompréhension de la consigne de la part des contributeurs de JeuxDeMots. Il est à noter que cette confusion entre hyponymie et méronymie – à laquelle nous nous retrouvons également confrontés dans le chapitre suivant, à la section 8.3.1 – a été observée chez Chaffin et Herrmann (1984) :

> The only disagreement with the a priori classification that cannot be explained in terms of defining properties of the relations was the sorting of the part-whole relation for places (*Germany/Hamburg*) with the class inclusion relations. This unexpected result appears to have been due to the subjects' confusing the part-whole relation of geographic inclusion with the class inclusion relation for geographic terms (*country/Russia*).

De tels exemples illustrent les limites que peuvent présenter les ressources obtenues par *crowdsourcing*. Elles nous incitent à prendre des précautions lors de la manipulation des données contenues dans JDM.

7.4.3 Rappel variable en fonction de la ressource

Le dernier cas de figure que nous avons observé est celui des hyperonymes dont le rappel varie le plus en fonction de la base de voisins. Nous avons précédemment observé un tel cas avec le nom *engin* (tableau 7.3), dont 87 % des

hyponymes étaient présents dans les VDW, 33 % dans les VDLM et 0 % dans les VDF. Les données que nous observons ici mettent en évidence l'influence de la nature du corpus sur le fonctionnement des mots en contexte et, par conséquent, sur les rapprochements générés par l'ADA. Cette démarche fait donc écho au travail que nous avons effectué sur les synonymes au chapitre 5.

Nous avons choisi d'étudier ici les 30 hyperonymes dont l'écart type est le plus élevé, c'est-à-dire ceux dont la proportion d'hyponymes captés varie le plus d'un corpus à l'autre (nous n'avons pas jugé pertinent d'analyser de façon distincte, parmi ces hyperonymes, ceux pour lesquels le rappel moyen est faible et ceux pour lesquels il est élevé). Ces données ont été rapportées au tableau 7.17.

L'analyse de ces hyperonymes montre que dans la quasi-totalité des cas, les hyponymes sont bien – voire très bien – captés par les VDW et les VDLM et ne le sont pas – ou très peu – par les VDF. Il est difficile de dire dans quelle mesure ce phénomène est à attribuer aux particularités que pourrait présenter le texte littéraire du corpus Frantext par rapport aux corpus encyclopédique de Wikipédia et journalistique du Monde, ou s'il s'agit plus simplement d'un effet lié à la différence de taille entre les trois corpus. Nous avons en effet vu que le corpus Frantext a une taille beaucoup plus réduite que celle des corpus Wikipédia et Le Monde, ce qui a une influence – défavorable – sur la précision des rapprochements produits par l'ADA.

On peut voir par exemple au tableau 7.18 que les VDW et les VDLM extraient exactement les mêmes hyponymes du nom *établissement*. Seul *bar* n'est pas extrait : dans les VDW, il a tendance à apparaître avec des prédicats comme *tenancier*_DE, *gérant*_DE, *serveur*_DANS, *écumer*_OBJ ou *pianiste*_DE, qui ne prennent pas *établissement* pour argument dans le corpus. On voit que dans les VDF, seul l'hyponyme *restaurant* est suffisamment substituable avec *établissement* pour qu'ils soient voisins. Malgré cela, l'observation des contextes communs à ces deux mots montre la fragilité de ce rapprochement. Nous avons rapporté au tableau 7.19 les 14 contextes communs à *établissement* et *restaurant*, triés par information mutuelle décroissante. On voit clairement que la plupart des contextes qui permettent de rapprocher ces deux mots sont extrêmement génériques. Ces contextes renvoient au fait que ces mots ont en commun d'être des lieux (*aller*_À, *quitter*_OBJ, *se trouver* _SUJ, etc.). Les modifieurs *petit* et *grand* sont encore plus vagues. Seuls les contextes *patron*_DE et *directeur*_DE semblent un tant soit peu discriminants (on voit que leur information mutuelle moyenne avec les deux voisins est la plus élevée). À titre de comparaison, les contextes qui permettent de rapprocher *établissement* et *restaurant* dans les VDW sont les suivants : *déjeuner*_DANS, *toilettes*_DE, *serveur*_DE, *huppé*, *fréquenté*, *branché*, *luxueux*, *fréquenter*_OBJ, *cuisine*_DE, etc. On voit que la quantité d'information véhi-

hyperonyme	rappel	écart type
jeu	0,67	0,47
policier	0,67	0,47
militaire	0,63	0,45
action	0,61	0,44
résultat	0,61	0,44
domaine	0,61	0,44
représentation	0,53	0,41
membre	0,44	0,4
culture	0,66	0,4
texte	0,72	0,39
argent	0,35	0,38
établissement	0,61	0,38
activité	0,52	0,37
région	0,63	0,37
passage	0,66	0,37
spectacle	0,62	0,36
individu	0,5	0,36
art	0,71	0,36
monument	0,61	0,36
énergie	0,47	0,35
œuvre	0,75	0,35
exposition	0,75	0,35
montagne	0,59	0,34
collège	0,67	0,34
ton	0,40	0,34
jeune	0,67	0,34
zone	0,48	0,34
connaissance	0,47	0,34
signe	0,68	0,34
peinture	0,71	0,33

TAB. 7.17 – Les 30 hyperonymes dont le rappel varie le plus.

établissement			
	VDW	VDLM	VDF
asile			✗
auberge			✗
bar	✗	✗	✗
cantine			✗
clinique			✗
collège	✓	✓	✗
hôpital	✓	✓	✗
hospice			✗
institution	✓	✓	✗
laboratoire	✓	✓	✗
lycée	✓	✓	✗
prison	✓	✓	✗
restaurant	✓	✓	✓
rappel	0,88	0,88	0,08

TAB. 7.18 – Repérage des hyponymes du nom *établissement* dans les trois bases de voisins.

culée par ces contextes est nettement plus élevée que dans ceux de Frantext. Les rapprochements générés à partir de l'analyse distributionnelle automatique de corpus volumineux comme Wikipédia ou Le Monde s'appuient donc, de fait, sur des vecteurs de contextes beaucoup plus riches que s'ils avaient été calculés sur un corpus de taille plus modeste comme Frantext. Il en résulte que la finesse de ces rapprochements est plus élevée dans les VDW et les VDLM que dans les VDF.

Il est toutefois intéressant de noter que parmi les hyperonymes du tableau 7.17 figure le nom *ton*. Nous avons montré à la section 5.5 une différence dans les synonymes de *ton* repérés dans les VDW et les VDLM. On constate que ce décalage s'observe également pour les hyponymes. On peut en effet voir, au tableau 7.20 que la majorité des hyponymes de *ton* qui sont extraits dans les VDW ne le sont pas dans les VDLM (ni dans les VDF). Le fait que *bleu*, *noir*, *rouge* et *vert* ne partagent pas les mêmes contextes que *ton* dans les VDLM semble ici aussi suggérer que ce mot n'est pas majoritairement employé pour désigner une nuance de couleur (on suppose que c'est parfois le cas étant donné que *ton* et *couleur* sont des VDLM).

mot	rel.	cat.	i. m.
patron	DE	N	13,327
directeur	DE	N	11,936
aller	DANS	V	10,089
trouver	DANS	V	8,53
faire	DANS	V	8,369
entrer	DANS	V	8,487
quitter	OBJ	V	7,057
porte	DE	N	7,708
aller	À	V	7,509
tenir	OBJ	V	5,619
petit	_	A	5,577
grand	_	A	5,801
faire	obj	V	4,139
se trouver	SUJ	V	7,982

TAB. 7.19 – Contextes communs à *établissement* et son hyponyme *restaurant* dans le corpus Frantext.

ton			
	VDW	VDLM	VDF
bleu	✓	✗	✗
couleur	✓	✓	✗
gris	✓		✗
jaune	✓		
noir	✓	✗	✗
note	✓	✗	✗
orange	✓		
rouge	✗	✗	✗
tonalité	✓	✓	✗
vert	✓	✗	
rappel	0,90	0,29	0

TAB. 7.20 – Repérage des hyponymes du nom *ton* dans les trois bases de voisins.

7.4.4 Conclusion

Dans ce chapitre, nous avons abordé la question des conditions qui influent sur le repérage de la relation d'hyperonymie par ADA. Pour ce faire, nous avons croisé nos bases de voisins avec des hypo/hyperonymes extraits de JDM que nous avons regroupés par hyperonyme. Cela nous a permis de décrire plusieurs types de fonctionnements qui peuvent favoriser ou défavoriser l'extraction d'un couples d'hypo/hyperonymes par l'ADA. Nous avons ainsi pu constater que parmi le groupe des hyperonymes qui manifestent le plus haut degré de substituabilité avec leurs hyponymes figurent les noms d'animés humains. L'étude des cas inverses – les hyperonymes qui ne partagent pas les mêmes contextes d'apparition que leurs hyponymes, et ce dans les trois bases de voisins – a mis au jour quatre phénomènes qui expliquent le décalage entre la distribution des hyperonymes et de leurs hyponymes :

- l'hyperonyme ou les hyponymes sont polysémiques et les acceptions qui sont réalisées dans le corpus ne sont pas celles entre lesquelles porte la relation d'hyperonymie ;
- certaines facettes des hyponymes émergent tellement dans le corpus que cela crée un décalage distributionnel avec les hyperonymes qui leurs sont attribués dans JDM (nous avons vu que ce cas de figure était assez proche de la polysémie) ;
- la relation d'hyperonymie porte sur des emplois *en mention* des hyponymes alors que ces derniers sont employés *en usage* dans les corpus ;
- la relation n'est pas hyperonymique.

Pour finir, nous avons étudié le cas des hyperonymes dont les hyponymes ont été repérés de façon fluctuante d'un corpus à l'autre. Les résultats se sont avérés peu concluants étant donné que la différence de taille entre les trois bases de voisins implique le plus souvent que les hyponymes sont le moins repérés dans les VDF. Une perspective de cette étude serait donc de la réitérer dans des conditions plus favorables, c'est-à-dire avec des bases distributionnelles plus équilibrées.

Chapitre 8

Étude des manifestations de la relation de méronymie dans une ressource distributionnelle

Sommaire

Dans ce chapitre, nous menons une étude qui a pour objet les manifestations de la relation de méronymie dans les bases distributionnelles. La relation de méronymie est intéressante à plusieurs titres : tout d'abord, elle constitue l'une des relations visées par les méthodes d'acquisition de ressources lexicales et terminologiques, au même titre que les relations plus souvent étudiées que sont l'hyperonymie et la synonymie. Ensuite, elle se décline en un ensemble varié de relations (CONSTITUANT/OBJET, ÉTAPE/ACTIVITÉ, MEMBRE/COLLECTION, etc.), ce qui offre un terrain d'observation particulièrement riche pour étudier les modalités d'application de l'analyse distributionnelle automatique (ADA). Enfin, contrairement aux autres relations

classiques, elle présente la particularité de pouvoir porter sur deux mots relevant chacun de classes sémantiques différentes : c'est le cas par exemple du couple *tête/enfant*, composé d'un premier nom qui renvoie à une partie du corps et d'un deuxième qui désigne un être humain, ou encore de *épée/métal*, qui relie un nom qui renvoie à un objet (une arme) à un autre qui désigne un matériau. De ce fait, puisque le principe de l'ADA est de rapprocher des mots qui présentent un certain degré de similarité sémantique, cette propriété des couples de méronymes semble défavorable à leur extraction automatique. Or, nous avons vu à la section 4.4 que les bases de voisins contenaient une quantité non négligeable de méronymes. Nous cherchons donc ici à mettre au jour les propriétés linguistiques de ces couples afin de mieux comprendre les raisons qui font que certains d'entre eux ont été captés par l'ADA alors que d'autres échappent au repérage.

La démarche que nous adoptons ici s'appuie sur un jeu de couples de méronymes issu de JeuxDeMots (JDM) que nous croisons avec les voisins de Wikipédia (VDW). Après avoir évoqué les propriétés principales de la relation de méronymie et des sous-relations qui la composent (8.1), nous présentons les couples que nous avons utilisés ainsi que la démarche qui a consisté à les croiser avec les voisins (8.2). Nous évoquons ensuite la phase d'annotation, qui a donné lieu à deux procédures successives (8.3), à savoir :

1. une sous-catégorisation manuelle des couples basée sur une typologie des relations de méronymie ;
2. une annotation semi-automatique de la classe sémantique des mots composant les couples de méronymes.

Ces annotations sont ensuite mises en relation avec le critère du voisinage distributionnel. Nous décrivons ainsi les propriétés qui font que l'ADA se montre particulèrement efficace pour extraire certains types de méronymes alors qu'elle est inadaptée pour d'autres (8.4).

8.1 La relation de méronymie : définition et typologie

La relation de méronymie est la relation qui s'établit entre une *partie* et son *tout*. Elle est asymétrique et sa réciproque, la relation entre un tout et l'une de ses parties, est l'holonymie. C'est une relation qui opère principalement entre deux noms, bien que Winston *et al.* (1987) proposent une relation FEATURE/ACTIVITY pour les couples désignant une étape dans un processus comme *paying/shopping*. La définition que donne Cruse (1986) de la méronymie est la suivante "X is a meronym of Y if and only if sentences

of the form *A Y has Xs / an X* and *An X is a part of a Y* are normal when the noun phrases *an X*, *a Y* are interpreted generically." Ainsi, *La main est une partie du bras* est vrai, et ce même s'il existe des bras dont la main a été coupée.

La relation de méronymie est prise en compte dans la construction des thésaurus et entre dans la structure des ontologies (Van Campenhoudt, 1996; Keet et Artale, 2008) (souvent sous le nom de relation *part of*). Elle se décline en plusieurs sous-relations. Winston *et al.* (1987) définissent six sous-types de méronymes en s'appuyant sur les trois critères suivants :

- la *fonctionnalité* : la partie a-t-elle une fonction vis-à-vis du tout ? Par exemple, *poignée* est fonctionnel vis-à-vis de *porte*, mais pas vis-à-vis de *maison* ;
- l'*homéomérité* : la partie et le tout sont-ils matériellement identiques ou différents (*tranche/gâteau* vs *arbre/forêt*) ?
- la *séparabilité* : la partie et le tout sont-ils séparables ? C'est le cas de *anse* et *tasse*, mais pas d'*acier* et *vélo*.

La combinaison de ces trois critères leur permet de dégager les relations suivantes, rapportées au tableau 8.1 :

- la relation ÉLÉMENT/OBJET porte sur des couples dans lesquels le méronyme est un élément saillant qui a un rôle fonctionnel vis-à-vis de l'holonyme : *réservoir/automobile*, *pince/crabe*, *boucle/ceinture* ;
- la relation MEMBRE/COLLECTION relie des couples où, pour un méronyme *m*, l'holonyme peut être désigné comme "un ensemble de *m*" : *automobile/convoi*, *mouton/troupeau* ou *arbre/forêt* ;
- la relation PORTION/MASSE porte sur des couples pour lesquels la partie représente un échantillon représentatif du tout : *goutte/vin*, *tranche/gâteau*, *centimètre/mètre* ;
- la relation CONSTITUANT/OBJET relie généralement des couples où le méronyme est une matière qui entre dans la composition de son tout : *plastique/dé*, *coton/drap*, *cuir/chaussure* ;
- la relation ÉTAPE/ACTIVITÉ relie un nom d'évènement à un autre évènement de plus grande ampleur dans lequel il se manifeste : *ravitaillement/marathon*, *charge/attaque* ou *discours/conférence* ;
- la relation LIEU/ZONE rapproche des noms de lieux qui se trouvent dans des lieux plus étendus : *salon/logement*, *forêt/montagne* ou *parc/quartier*.

En marge de cette première série, Winston *et al.* (1987) décrivent une série de relations d'inclusion qui peuvent – dans une certaine mesure – s'apparenter à de la méronymie mais qui n'en sont pas. Ces relations *pseudo-méronymiques* sont les suivantes :

- l'inclusion topologique : l'holonyme est un contenant (*prisonnier/cellule*), une zone (*Berlin Ouest/Allemagne de l'Est*) ou exprime une durée tem-

relation	exemple	critères		
		fonct.	homéo.	sépar.
ÉLÉMENT/OBJET	*anse/tasse*	+	-	+
MEMBRE/COLLECTION	*arbre/forêt*	-	-	+
PORTION/MASSE	*tranche/gâteau*	-	+	+
CONSTITUANT/OBJET	*acier/vélo*	-	-	-
ÉTAPE/ACTIVITÉ	*payer/magasiner*	+	-	-
LIEU/ZONE	*oasis/désert*	-	+	-

TAB. 8.1 – Sous-types de la relation de méronymie définis par Winston *et al.* (1987).

porelle (*réunion/matin*) ;
- l'inclusion de classe : il s'agit ici de la relation d'hyponymie (*rose/fleur*, *peur/émotion*, etc.). Le fait de marquer la différence entre cette relation et la méronymie a certainement été motivé par les travaux de Chaffin et Herrmann (1984), qui observent une tendance chez les locuteurs à les confondre (cf. section 8.1) ;
- la relation d'attribution : il s'agit d'une relation de type modifieur entre un mot et un adjectif (*tour/haute*, *blague/drôle*, etc.) ;
- la relation d'attachement : elle porte sur deux objets physiquement attachés l'un à l'autre (*boucle d'oreille/oreille*) ;
- la relation d'appartenance : elle relie des mots comme *millionnaire* et *argent* ou *auteur* et *copyright* et peut être confondue avec la méronymie à cause de l'ambiguïté du patron *X a Y*, qui peut exprimer l'appartenance (*Camille a un vélo* vs *Un vélo a des roues*).

La différence entre certaines de ces relations et la méronymie *stricto sensu* est parfois assez fine. Nous verrons à la section 8.3 que beaucoup des paires annotées relèvent de l'une de ces relations pseudo-méronymiques.

8.2 Croiser les VDW et un jeu de méronymes

À l'instar des études que nous avons menées dans les chapitres 5 et 7, nous sélectionnons nos données d'observation par le croisement des voisins et d'un jeu de relations lexicales contenues dans un lexique externe. Nous ne revenons pas sur la description des VDW et JDM, cette dernière ayant été faite – respectivement – aux sections 3.2 et 4.3.3.1.

Dans JDM, les couples de méronymes sont ajoutés au réseau par des joueurs à qui le programme a présenté l'une ou l'autre des consignes sui-

vantes :

> Donner des PARTIES du terme suivant : (une partie est une composante de l'objet, par exemple : *moteur*, *roue*, etc. pour *voiture* – ou encore *couverture*, *pages*, *chapitre* etc. pour *livre*) ;
>
> Donner des TOUT du terme suivant : (le tout est ce qui englobe/contient/possède la partie, par exemple : *corps*, *bras*, etc. pour *coude* – ou encore *banque* pour *guichet*) ;

Comme nous l'avons vu à la section 4.3.3.1, les couples de méronymes ne sont pas symétrisés. Ils apparaissent dans le réseau de façon distincte dans les relations HAS PART et HOLO. On compte respectivement 12 965 et 9872 occurrences de ces relations dans la version du réseau qui était disponible au moment où nous avons réalisé cette étude, à savoir celle du 10 mai 2011[1]. Nous avons donc récupéré de JDM ces 22 837 couples de noms entretenant une relation de méronymie ou d'holonymie.

Ici encore, ces couples ont subi plusieurs traitements avant d'être comparés aux voisins. Nous ne revenons pas sur ces modifications, qui sont les mêmes que celles qui ont été appliquées sur les couples d'hypo/hyperonymes aux sections 7.2.2 et 7.3.2. La seule différence se situe au niveau du Rprod utilisé. En effet, dans cette étude, seuls les couples pour lesquels le Rprod est supérieur ou égal à 0,60 ont été conservés. Ce seuil, beaucoup plus restrictif que celui de 0,23 que nous avons fixé lors de notre étude de la relation de la synonymie – chronologiquement ultérieure à la présente étude – dans le chapitre 5, a été fixé de façon à ne filtrer qu'une quantité réduite de couples, dans l'optique d'une révision manuelle des données. Ainsi, dans un couple de méronymes, un mot ne pourra pas avoir une productivité 40 % plus élevée ou plus basse que l'autre mot. Pour rappel, cette manipulation consiste à atténuer la tendance de la mesure de similarité distributionnelle à favoriser le repérage des couples composés de mots qui ont des fréquences – et des productivités – comparables (cf. préambule méthodologique).

La base obtenue (désormais $\text{JDM}_{méro}$) compte 1520 paires. Le croisement de ces données avec les VDW montre que 55 % des couples (soit 829) sont captés par l'ADA. L'étape suivante consiste à annoter ces couples afin de mettre au jour des différences entre les couples de méronymes qui ont été captés par l'ADA et ceux qui ne l'ont pas été.

[1] Comme nous l'avons indiqué quand nous avons présenté la ressource à la section 4.3.3.1, le nombre d'occurrences de chacune de ces relations était de 23 641 et de 13 488 dans la version de JDM du 5 février 2013.

8.3 Phase d'annotation

La phase d'annotation doit permettre de prendre en compte parmi les couples de JDM$_{méro}$ la diversité des sous-relations à travers lesquelles se décline la relation de méronymie. Le but de notre étude étant de caractériser les différentes façons dont se manifeste la relation de méronymie parmi les voisins distributionnels, nous procédons à une distinction des différents types de couples contenus dans la base JDM$_{méro}$. Nous nous appuyons dans un premier temps sur la typologie de Winston *et al.* (1987) décrite à la section 6.1.1, pour tester l'hypothèse selon laquelle les couples ÉLÉMENT/OBJET, PORTION/MASSE, etc. se manifestent selon des conditions différentes parmi les voisins. Les difficultés rencontrées lors de cette annotation nous amènent, dans un deuxième temps, à nous orienter vers une démarche inspirée de Murphy (2003) visant à identifier la classe sémantique des mots qui entrent dans une relation de méronymie.

8.3.1 Typologie de Winston *et al.* (1987)

Cette annotation étant entièrement manuelle, nous avons dans un premier temps fait le choix de n'annoter qu'un sous-ensemble des 1520 couples sélectionnés à la section précédente. Le Rprod nous a permis de sélectionner cet échantillon : nous avons ainsi choisi d'annoter les 481 couples dont le Rprod est supérieur ou égal à 0,85. L'annotation a été réalisée par un seul individu – nous-mêmes –, ce qui ne permet pas de calculer un accord inter-annotateur qui viendrait attester de la stabilité des annotations. La raison en est qu'il s'agissait en premier lieu d'estimer la faisabilité de la tâche d'annotation des couples de méronymes (pour éventuellement mobiliser des annotateurs par la suite). Les résultats nous ont convaincu de ne pas aller plus loin dans cette direction.

En effet, la difficulté qu'il y a à annoter des couples de mots hors contexte – que nous avons commentée aux sections 4.2.2 et 6.3.2 – est ici accentuée par le fait qu'un grand nombre de couples portent des relations dont le caractère méronymique pourrait être discuté. De ce fait, les résultats de l'annotation, dont nous avons rapporté un extrait dans le tableau 8.2, sont extrêmement contestables. Par exemple, si on voit assez clairement comment *réservoir*/*voiture* entretiennent une relation ÉLÉMENT/OBJET, c'est plus discutable pour PISTE/ENQUÊTE, qui, à notre avis, n'entre pas tout à fait dans la catégorie ÉTAPE/ACTIVITÉ. On aurait plutôt affaire à une relation de type scénario. Dans le cas du couple *arrivée*/*circuit*, annoté LIEU/ZONE, *arrivée* pourrait très bien être interprété comme un événement ayant lieu sur un circuit et non comme une zone du circuit.

relation identifiée	couple	voisins ?
ÉLÉMENT/OBJET	*oreille/loup*	✓
	piste/enquête	✓
	réservoir/automobile	
CONSTITUANT/OBJET	*métal/couronne*	
	botte/caoutchouc	
	coton/drap	✓
LIEU/ZONE	*salon/logement*	✓
	accueil/immeuble	
	arrivée/circuit	
MEMBRE/COLLECTION	*automobile/convoi*	✓
	plante/jardin	✓
	mouton/troupeau	✓
ÉTAPE/ACTIVITÉ	*ravitaillement/marathon*	
	victoire/bataille	✓
	discours/conférence	✓
?	*hôpital/santé*	
	héros/légende	✓
	électricité/four	
INCLUSION TOPOLOGIQUE	*animal/forêt*	✓
	docteur/hôpital	
	musicien/concert	✓
CLASSE	*femme/personne*	✓
	coq/volaille	
	collier/bijou	
SYNONYMIE	*blague/gag*	
	repos/sommeil	✓
	fin/terme	✓
APPARTENANCE	*coureur/maillot*	
	indien/plume	
	apôtre/robe	

TAB. 8.2 – Exemples de couples annotés selon la typologie de Winston *et al.* (1987).

	relation	fréq.	proportion
relations méronymiques	ÉLÉMENT/OBJET	177	36,8 %
	CONSTITUANT/OBJET	35	7,3 %
	LIEU/ZONE	34	7,1 %
	MEMBRE/COLLECTION	15	3,1 %
	ÉTAPE/ACTIVITÉ	6	1,2 %
autres relations	?	151	31,4 %
	INCLUSION TOPOLOGIQUE	31	6,4 %
	CLASSE	18	3,7 %
	SYNONYMIE	11	2,3 %
	APPARTENANCE	3	0,6 %

TAB. 8.3 – Résultats de l'annotation basée sur la typologie de Winston *et al.* (1987).

Certains cas sont tellement problématiques que nous avons créé une étiquette " ?" que nous avons attribuée aux couples comme *hôpital/santé*, *héros/légende* ou *four/électricité* pour lesquels la relation relève davantage de l'association libre que de la méronymie. Il est difficile d'expliquer les raisons qui font que les joueurs de JeuxDeMots produisent ce type de paires. On peut avancer l'hypothèse d'une mécompréhension de la consigne ou d'une conception particulièrement lâche de la relation partie-tout (on peut en effet considérer que le héros est le protagoniste d'une légende, donc qu'il en est une partie). Il est également probable que les mécanismes de jeu de JeuxDeMots poussent les participants à produire des relations comme *cœur/particulier*, *goutte/orage* ou *tronc/malade* qui, sans être erronées – un particulier a un cœur, une goutte est une partie d'un orage et un malade a un tronc – paraissent assez peu naturelles, voire forcées. Ainsi, un joueur à qui il est demandé de donner des parties de *particulier* n'a d'autre choix que de soumettre des parties du corps puisque le sens "personne privée" de *particulier* apparaît difficilement interprétable comme un holonyme.

Cette annotation nous a malgré tout permis d'avoir une vue d'ensemble des types de méronymes contenus dans notre échantillon. La répartition des paires dans les catégories a été rapportée au tableau 8.3. Les relations décrites comme méronymiques dans la typologie figurent dans la partie haute du tableau, les relations pseudo-méronymiques apparaissent en bas. On peut constater que c'est la relation ÉLÉMENT/OBJET qui prévaut, puisqu'elle représente 36,8 % des couples de l'échantillon. Les 4 autres relations méronymiques – au sens strict – sont nettement minoritaires.

Dans la partie basse du tableau, les relations pseudo-méronymiques re-

présentent 44 % des couples annotés. Les relations telles que l'inclusion topologique, la relation de classe, la synonymie et l'appartenance représentent 13 % des couples. La relation qui prévaut est la relation identifiée par un point d'interrogation, que nous avons évoquée plus haut. Elle concerne 31,4 % des couples de l'échantillon. Certains des couples de cette catégorie passent le test de l'insertion dans un patron de type *X a Y* (*chauffeur*/*taxi*, *carte*/*couleur*, *billet*/*montant*). Beaucoup de ces paires sont en fait constituées de noms exprimant des concepts abstraits pour lesquels les critères de fonctionnalité, d'homéomérité et de séparabilité sont difficilement applicables comme dans *calcul*/*chiffre* ou *lumière*/*univers*. Sur ce point, il est important de rappeler que les relations ont été produites par des locuteurs *tout-venant* qui avaient pour seule consigne de "donner des TOUT/PARTIES" des mots-cibles qui leur étaient proposés. Malgré le caractère périphérique d'une partie des relations, nous prenons comme objet d'étude le jeu de couples dans son ensemble, dans la mesure où ils ont été produits par des locuteurs qui les ont perçus comme relevant de la relation partie/tout.

Le bilan que l'on peut tirer de cette première annotation est que la typologie montre ses limites lorsqu'elle est confrontée aux données de JDM, pour deux raisons :

- une seule relation – ÉLÉMENT/OBJET – concentre près de 66 % des couples considérés comme relevant strictement de la méronymie. Cette classe apparaît manifestement comme trop englobante dans la mesure où elle porte sur des couples de nature hétérogène ;
- 31,4 % des couples ne relèvent pas d'une des relations définies dans la typologie de référence, même en l'augmentant avec la série des relations pseudo-méronymiques.

Nous avons donc décidé de délaisser une typologie préétablie et d'adopter une approche *bottom-up* : nous nous focalisons cette fois sur le sens des mots composant les paires de méronymes afin de faire émerger des combinaisons de classes sémantiques.

8.3.2 Annotation en classes sémantiques

La deuxième procédure d'annotation consiste à attribuer une classe sémantique à chaque mot des couples de méronymes, afin de mettre au jour de nouvelles configurations distributionnelles. Elle s'inspire du point de vue de Murphy (2003), qui rejette l'idée selon laquelle il existerait plusieurs déclinaisons de la méronymie et qui considère que la seule chose qui change entre les différents sous-types est la nature des mots sur lesquels porte la relation.

Les couples que nous avons utilisés sont ceux de la base $\text{JDM}_{méro}$: la méthode d'annotation étant cette fois semi-automatisée, nous avons utilisé

un ensemble plus élevé de paires que dans la section précédente. Suite aux résultats obtenus lors de l'annotation effectuée à partir de la typologie de Winston *et al.* (1987), nous avons choisi de retirer manuellement les couples d'hyperonymes et de synonymes. JDM$_{méro}$ compte désormais 1334 paires dont 53 % (711) sont détectées par l'ADA (contre 1520 paires dont 55 % de voisins dans sa version précédente).

En guise de classe sémantique, nous avons associé chaque mot à l'un de ses hyperonymes *de haut niveau* dans WordNet. Cette démarche a, dans un premier temps, consisté à traduire le lexique de JDM$_{méro}$ en anglais[2]. Nous avons ensuite associé chaque mot à l'ensemble des hyperonymes de sa traduction anglaise dans le réseau, et ce pour chacune de ses acceptions recensées dans WordNet. Ainsi, *église* est associé aux quatre chemins suivants :

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>EVENT>HUMAN_ACTIVITY>ACTIVITY>CEREMONY
ENTITY>ABSTRACT_ENTITY>ABSTRACTION>SOCIAL_GROUP>ORGANIZATION>INSTITUTION>RELIGION
ENTITY>ABSTRACT_ENTITY>ABSTRACTION>SOCIAL_GROUP>GATHERING>BODY
ENTITY>PHYSICAL_ENTITY>PHYSICAL_OBJECT>WHOLE>ARTIFACT>STRUCTURE>EDIFICE>PLACE_OF_WORSHIP

L'étape suivante consiste à procéder à un *élagage* de l'arborescence. Par défaut, la granularité de WordNet est bien trop fine pour nous permettre d'obtenir des classes de taille satisfaisante (par exemple, dans nos données, *église* est le seul mot à figurer en position hyponyme de CEREMONY). L'élagage vise à obtenir une arborescence moins complexe. Ainsi, nous avons choisi de couper les noms abstraits (hyponymes de ABSTRACT_ENTITY) au troisième niveau de profondeur et les noms concrets (hyponymes de PHYSICAL_ENTITY) au cinquième. Ce choix se justifie par un nombre plus important de noms concrets dans nos données. Dans le cas de *église*, cela entraîne la disparition de la nuance entre les deux acceptions du mot en tant que groupe social :

ENTITY>ABSTRACT_ENTITY>ABSTRACTION>EVENT
ENTITY>ABSTRACT_ENTITY>ABSTRACTION>SOCIAL_GROUP
ENTITY>PHYSICAL_ENTITY>PHYSICAL_OBJECT>WHOLE>ARTIFACT>STRUCTURE

Les mots ainsi annotés sont ensuite désambiguïsés manuellement en fonction du mot avec lequel ils entretiennent une relation de méronymie. Dans l'exemple précédent, cette démarche consiste à associer *église* à un type de bâtiment (STRUCTURE) dans le couple *église/village* et à un groupe social (SOCIAL_GROUP) dans le couple *fidèle/église*. Toujours dans la même optique, nous avons procédé à différents ajustements consistant à opérer des regroupements entre certaines classes de mots. Cette étape s'est faite de façon empirique en fonction notamment du nombre d'éléments contenus dans chaque catégorie : par exemple, les éléments appartenant à des catégories de moins de 10 membres ont été systématiquement déplacés sous l'hyperonyme

[2]Cette étape a été facilitée par l'utilisation de Google Traduction (`http://translate.google.fr/`). Les traductions ont ensuite été vérifiées manuellement.

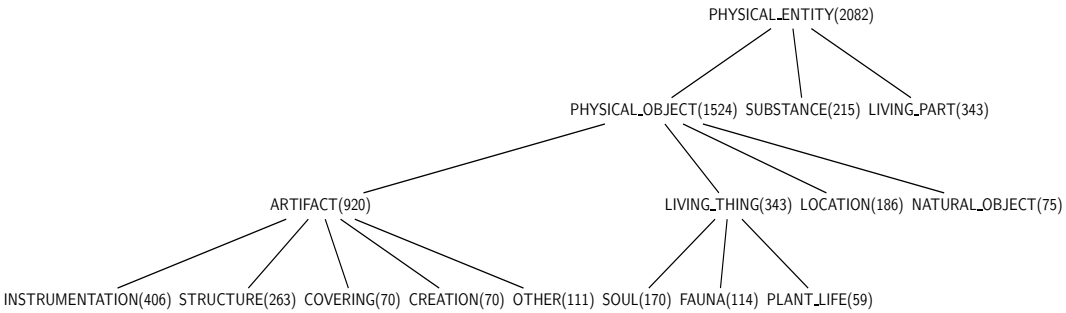


FIG. 8.1 – Répartition des mots de JDM$_{méro}$ dans la classe PHYSICAL_ENTITY.

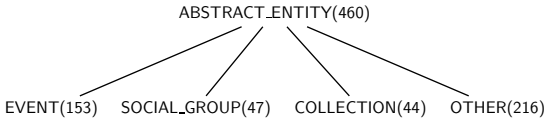


FIG. 8.2 – Répartition des mots de JDM$_{méro}$ dans la classe ABSTRACT_ENTITY.

de niveau supérieur. Dans l'exemple ci-dessous, le premier chemin correspond à celui de *doigt*, le deuxième à celui de *nez* :

ENTITY>PHYSICAL_ENTITY>THING>PIECE>BODY_PART>EXTERNAL_BODY_PART>MEMBER>DIGIT
ENTITY>PHYSICAL_ENTITY>THING>PIECE>BODY_PART>ORGAN>SENSORY_RECEPTOR>CHEMORECEPTOR

Après regroupement, les deux mots se retrouvent au même niveau dans la hiérarchie : ils sont directement subordonnés à BODY_PART. La répartition finale des mots de JDM$_{méro}$ après désambiguïsation et élagage de la hiérarchie a été rapportée à la figure 8.1 pour les noms concrets et à la figure 8.2 pour les noms abstraits.

Sur ces figures, on constate clairement que la profondeur de la hiérarchie varie selon les classes. La classe ABSTRACT_ENTITY n'a qu'un niveau de profondeur. Elle se divise en quatre classes : les événements (EVENT : *exposition*, *procès*), les groupes sociaux (SOCIAL_GROUP : *peuple*, *famille*), les collections (COLLECTION : *flotte*, *galaxie*) et *autres* (OTHER : *trou*, *valeur*). La catégorie *autres* est un ajout de notre part, elle contient des mots appartenant à des classes comme les jours de la semaine, les unités monétaires ou les notes de musique qui contiennent trop peu de membres pour avoir une existence autonome dans notre classification. La classe des noms concrets regroupe un nombre de noms beaucoup plus important que la classe des noms abstraits (2082 contre 460). Elle est structurée de façon plus com-

plexe et possède trois niveaux de profondeur. Le premier niveau distingue les parties du corps (LIVING_PART, qui regroupe en fait les parties de corps humain, animal, et les parties de végétaux : *langue*, *patte*), les substances ou matières (SUBSTANCE : *fer*, *laine*), et les objets physiques. Cette dernière classe, la plus volumineuse, regroupe les objets naturels (NATURAL_OBJECT : *torrent*, *volcan*), les lieux (LOCATION : *grotte*, *quartier*), les entités vivantes (LIVING_THING) et les artéfacts (ARTIFACT). Ces deux dernières classes possèdent enfin un dernier niveau de profondeur. La classe des entités vivantes se subdivise en trois sous-classes regroupant les noms se rapportant à des humains (SOUL : *joueur*, *pompier*), des animaux (FAUNA : *canard*, *requin*) et des végétaux (PLANT_LIFE : *olive*, *rose*). La classe des artéfacts comprend les noms d'instruments (INSTRUMENT : *lampe*, *pneu*), de bâtiments (STRUCTURE : *lycée*, *magasin*), de vêtements (COVERING : *cape*, *chapeau*), les créations (CREATION : cette classe englobe les productions littéraires, artistiques comme *fresque* ou *roman*) et une catégorie *autres* (OTHER : de la même façon que la catégorie éponyme dans la classe des entités abstraites, elle regroupe des objets de nature hétérogène comme *brique* ou *savon*).

L'abandon d'une typologie préétablie au profit de classes sémantiques va nous permettre de mener des analyses plus précises : dans la section suivante, nous analysons les propriétés distributionnelles des couples de classes les plus fréquents dans notre jeu de méronymes.

8.4 Analyse des couples

À ce stade de l'étude, nous nous intéressons à deux propriétés des 1334 couples de la base JDM$_{méro}$:

- chaque couple a été catégorisé selon qu'il a été capté par l'ADA ou non ;
- chaque membre de chaque couple de méronymes est associé a une étiquette sémantique qui lui est propre.

Nous croisons à présent ces deux aspects afin de mettre au jour des couples de classes de mots et d'expliquer pourquoi certains sont mieux captés par l'ADA que d'autres. Nous avons rapporté au tableau 8.4 les couples de classes représentés par au moins 10 paires de méronymes dans la base (nous avons supprimé la classe OTHER à cause du caractère hétérogène des relations qu'elle englobe). Ils sont classés par ordre de proportion de voisins décroissante. On peut constater qu'il y a de fortes disparités entre les différentes combinaisons : alors que 96,2 % des couples dont le méronyme est un humain et l'holonyme un groupe social sont extraits comme des voisins par l'ADA, cela n'est vrai que de 14,6 % des couples dont le méronyme est une partie du corps et l'holo-

classe méro.	classe holo.	nb. de couples	% voisins
SOUL	SOCIAL_GROUP	54	96,2
LOCATION	LOCATION	19	90,0
EVENT	EVENT	29	85,3
STRUCTURE	LOCATION	94	82,6
STRUCTURE	STRUCTURE	12	82,5
SOUL	STRUCTURE	25	66,7
INSTRUMENTATION	STRUCTURE	11	51,9
SUBSTANCE	SUBSTANCE	27	51,2
INSTRUMENTATION	INSTRUMENTATION	82	41,5
LIVING_PART	PLANT_LIFE	76	41,2
LIVING_PART	LIVING_PART	62	33,9
SUBSTANCE	INSTRUMENTATION	17	33,3
LIVING_PART	SOUL	54	18,4
LIVING_PART	FAUNA	41	14,6

TAB. 8.4 – Couples de classes les plus fréquents dans JDM$_{méro}$.

nyme un animal. Cela signifie que les 54 couples SOUL/SOCIAL_GROUP sont composés de mots qui ont une forte tendance à être substituables alors que c'est beaucoup moins le cas pour les 41 couples LIVING_PART/FAUNA. Dans cette section, nous nous focalisons sur l'observation des propriétés distributionnelles de ces différents types de classes afin d'expliquer pourquoi certaines sont plus compatibles que d'autres.

Les différentes combinaisons de classes sont analysées en deux temps. La section 8.4.1 est consacrée à l'étude des couples constitués de deux mots appartenant à des classes identiques – couples dits *homogènes*. Les couples constitués de deux mots relevant de deux classes différentes – couples *hétérogènes* – sont analysés à la section 8.4.2. Ce découpage est motivé par le fait que, contrairement aux autres relations classiques, la méronymie possède la particularité de pouvoir associer deux mots qui ont des natures sémantiques différentes (*vache/troupeau*, *métal/épée*). Or, on sait que l'ADA basée sur l'analyse des contextes syntaxiques présente une tendance à rapprocher des mots qui sont sémantiquement similaires. Les données dont nous disposons nous donnent la possibilité de mettre au jour les conditions dans lesquelles se principe se vérifie ou ne se vérifie pas.

8.4.1 Couples homogènes

Parmi les 14 couples de classes rapportés au tableau 8.4, 6 sont homogènes. Leur proportion de voisins moyenne est de 64,1 %, ce qui est un peu plus élevé que celle des couples hétérogènes, qui est de 50,6 %.

8.4.1.1 Les classes les mieux repérées

Les couples composés de deux éléments appartenant aux classes LOCATION, EVENT ou STRUCTURE sont repérés par l'ADA dans des proportions allant de 82,5 % à 90 %. Les couples dont les deux membres appartiennent à la classe LOCATION expriment une relation entre deux lieux, l'un étant localisé dans un second de taille supérieure (*allemagne/europe*, *commune/ville*, *place/village*). Ce sont les couples homogènes qui sont le mieux repérés par l'ADA. Les mots qui les composent partagent la propriété d'exprimer des entités localisées spatialement. De fait, ils partagent des contextes comme la position objet de verbes de localisation – *(se) situer*, *se trouver* – *via* des prépositions complexes comme AU SUD DE ou AU CENTRE DE, ou encore la position complément du nom, quand le nom exprime un point cardinal (NORD DE, SUD DE, etc.). En plus de partager ce faisceau de contextes, les mots exprimant des lieux se distinguent par des contextes spécifiques qui permettent de distinguer des sous-classes le lieux. Par exemple, beaucoup des couples de lieux expriment différents niveaux de subdivisions administratives (*commune/canton*, *village/département*, etc.). Ils partagent des contextes comme *administration*_DE, *communauté*_DE, *population*_DE ou *territoire*_DE. De la même façon, l'analyse des contextes du couple *propriété/parc* montre qu'ils ont été rapprochés à la fois grâce au fait qu'ils sont des objets localisés dans l'espace (*limite*_DE, *s'étendre*_SUR, *superficie*_DE), mais aussi parce qu'il apparaissent comme des biens que l'on peut posséder (*revendre* _OBJ, *acheter*_OBJ, *gérer* _OBJ). Ces contextes spécifiques viennent renforcer la proximité distributionnelle entre les différents sous-ensembles de la classe des noms de lieux.

Le cas des couples d'événements est assez similaire si ce n'est que les mots expriment des valeurs temporelles et non plus spatiales : l'événement méronyme prend place dans un processus de plus grande ampleur exprimé par le second membre de la paire (*bataille/campagne*, *départ/course*, *victoire/combat*). Les noms exprimant une durée s'emploient dans des contextes comme *avoir lieu*_SUJ, *prendre fin*_SUJ et *se terminer*_SUJ, par l'intermédiaire de prépositions comme LORS DE, AU COURS DE, etc. Ici aussi, certains types d'événements se distinguent du fait, par exemple, que certains ont un aspect duratif alors que d'autres sont plus ponctuels (*mission* vs *vic-*

toire). Comme ça été le cas pour les noms de lieux, les contextes exprimant la localisation temporelle d'un événement sont associés à d'autres contextes exprimant des caractéristiques liées au sous-type d'événement.

Enfin, la classe STRUCTURE relie des noms de bâtiments ou de parties de bâtiments qui se situent au sein d'un autre bâtiment (*tour/château*, *hall/immeuble*, *salle/lycée*). Il semblerait que la distribution de la classe des bâtiments ait une distribution moins bien circonscrite que celle des lieux et des événements. En effet, l'étude des paires de cette catégorie ne fait apparaître que peu de contextes transversaux, qui s'appliquent à l'ensemble des mots appartenant à la classe des bâtiments. Le contexte *construire*_OBJ en est un : il peut virtuellement s'appliquer à tout type de bâtiment mais ne permet pas, par exemple, de rapprocher la paire *salon/appartement*. De la même façon, le contexte *habiter*_OBJ est assez répandu mais ne s'applique, par définition, qu'aux structures destinées à être habitables (ce contexte n'apparaît pas dans les contextes communs de la paire de voisins *pièce/musée*, par exemple). Ainsi, la classe des bâtiments apparaît de façon assez floue, dans la mesure où l'emploi qui est fait des noms de bâtiments dans le corpus met l'accent sur leur aspect fonctionnel. Il semblerait que les classes qui émergent se situent à un niveau de granularité inférieur. Par exemple, les couples *appartement/immeuble* et *chambre/hôtel* possèdent en commun des contextes comme *habiter*_OBJ, *louer*_OBJ ou *se installer*_DANS. Ces contextes définissent un type de bâtiment bien particulier, à savoir les bâtiments destinés au logement. De la même façon, les couples *tour/château* et *fortification/fort* partagent des contextes comme *protéger*_OBJ, *détruire*_OBJ ou *attaquer*_OBJ qui permettraient de dessiner les contours de la classe des bâtiments militaires.

Ainsi, les mots qui appartiennent à ces trois types de couples de classes sont particulièrement bien repérés par l'ADA du fait que les propriétés sémantiques qu'ils partagent se répercutent sur le plan distributionnel. Nous avons vu que les distributions des couples de lieux et d'événements se caractérisaient par un ensemble de contextes compatibles avec la plupart des mots appartenant à chacune de ces classes. Ce constat se vérifie dans une moindre mesure sur la classe des bâtiments, pour laquelle nous avons vu que les classes qui émergeaient se situaient à un niveau plus fin.

8.4.1.2 Classes repérées en quantités moindres

Les couples composés de deux éléments appartenant aux classes SUBSTANCE, INSTRUMENTATION ou LIVING_PART sont captés par l'ADA dans des proportions allant seulement de 33,9 % à 51,2 %. Nous avons donc affaire à des couples de mots qui, tout en possédant la même étiquette sémantique, se caractérisent par des propriétés distributionnelles différentes.

Dans le cas de la classe SUBSTANCE, les mots reliés désignent deux substances ou matières (au sens large) dont l'une entre dans la composition de l'autre : *carbone/diamant*, *crème/beurre*, *éthanol/rhum*. Nous identifions deux phénomènes expliquant la raison pour laquelle ces couples de mots sont mal repérés par l'ADA. Le premier est que leurs membres ne sont pas forcément employés comme des substances dans le corpus. *Rhum*, par exemple, apparaît comme un produit fini et non comme un ingrédient (sauf dans le contexte *baba*_À). Le second est que, même dans les – rares – cas où les deux mots sont employés comme des composants, ils n'entrent pas forcément dans la composition du même type d'objets : pour le couple *carbone/diamant*, les contextes comme *collier*_DE sont incompatibles avec *carbone*. Un couple comme *crème/beurre* fait exception à la règle. *Crème* et *beurre* ont été détectés comme voisins, ils partagent les contextes *mélanger*_OBJ, *incorporer*_OBJ, *verser*_OBJ, etc. Ces deux mots ont en commun qu'ils apparaissent comme des ingrédients de cuisine. Dans le cas de *carbone/diamant*, les contextes se recoupent moins dans la mesure où on a affaire, d'un côté, à un élément chimique et, de l'autre, à un minéral. Cette différence sémantique semble suffisamment importante pour qu'elle soit perceptible au niveau des distributions respectives de ces deux mots et qu'ils ne soient donc pas repérés comme des voisins.

Les couples dont les deux membres appartiennent à la classe INSTRUMENTATION sont extraits par l'ADA à hauteur de 41,5 %. La notion d'*instrument* est à prendre au sens large, et les couples appartenant à cette classe expriment une relation où un élément fait partie d'un dispositif ou un système de plus grande ampleur : *écran/ordinateur*, *pédale/bicyclette*, *pneu/autobus*. Dans la plupart des cas, les distributions entre les deux mots sont trop éloignées pour que l'analyse permette de les rapprocher. Par exemple, les contextes dans lesquels apparaît le méronyme *réservoir* (*volume*_DE, *servir*_DE, *placer*_OBJ) diffèrent complètement de ceux dans lesquels apparaissent ses holonymes *automobile* et *moto* (*accident*_DE, *conduire*_OBJ, *modèle*_DE). Le cas du méronyme *moteur*, en revanche, illustre une situation où la distribution du méronyme et de l'holonyme se recoupent : les 7 paires dans lesquelles il prend place sont toutes repérées par les voisins. Il apparaît en position méronyme de *avion*, *bateau*, *machine*, *navire*, *train*, *véhicule* et *voiture*. L'analyse des contextes communs fait apparaître une certaine symbiose entre le moteur et la machine qu'il équipe, dans la mesure où ils partagent un éventail de contextes relativement étendu comme *panne*_DE, *bruit*_DE, *consommer*_SUJ, *puissance*_DE, *s'arrêter*_SUJ, *fonctionner*_SUJ, etc. On pourrait analyser certains de ces contextes comme des cas de métonymie : le bruit produit par l'avion est en fait le bruit du moteur, de même que la puissance de la voiture est, encore une fois, celle de son moteur.

Les couples de mots appartenant tous deux à la classe LIVING_PART sont les couples homogènes les moins bien identifiés par l'ADA. Ils relient deux parties du corps (corps humain, animal ou partie d'un végétal), dont l'une est elle-même une partie de l'autre : *chair/doigt*, *muscle/bras*, *peau/visage*. Une des raisons expliquant les différences de distribution parmi les parties du corps est que, dans la plupart des cas, on a affaire à des sous-classes de parties du corps dont les fonctionnements diffèrent radicalement. Ainsi, le fait que les couples *nerf/jambe* ou *nerf/doigt* ne soient pas captés s'explique par le fait que *jambe* et *doigt* sont des membres du corps. Ils peuvent par conséquent apparaître en position objet de verbes comme *lever*, *croiser* ou *replier*, soit autant de contextes dans lesquels ne peut pas apparaître *nerf*.

Ainsi, nous pouvons conclure de l'analyse de ces trois types de couples que les catégories *substance*, *instrumentation* et *living_part* s'avèrent peu pertinentes du point de vue distributionnel. Elle sont constituées de mots dont les distributions sont particulièrement dissemblables. Ainsi, si l'on postule *a priori* l'existence d'une classe sémantique des parties du corps, l'analyse du corpus montre que les mots *jambe*, *bras*, *doigt*, etc. entrent en fait dans un paradigme différent de celui de *veine*, *nerf* ou *os*. Il y a donc un décalage entre les classes sémantiques que l'on pourrait dégager intuitivement et les classes distributionnelles qui émergent de l'analyse du texte.

8.4.2 Couples hétérogènes

Nous avons auparavant évoqué la tendance qu'a l'ADA à faire émerger des rapprochements relevant de la similarité sémantique, c'est-à-dire des mots qui sont "le même genre de choses" (van der Plas, 2008). De ce fait, on pouvait s'attendre à ne pas trouver de couples hétérogènes parmi les voisins distributionnels. Les résultats montrent toutefois que certains couples de catégories hétérogènes sont quasi-intégralement repérés par l'ADA.

C'est notamment le cas des couples de mots dont le méronyme appartient à la classe des humains (SOUL) et l'holonyme à celle des groupes sociaux : *capitaine/marine*, *fils/famille*, *musicien/orchestre*. Ces couples sont repérés à 96,2 %. Cela s'explique par le fait que les mots de la classe SOCIAL_GROUP ont des distributions similaires à ceux de la classe SOUL. Ils partagent par exemple la propriété d'apparaître en position sujet des verbes d'actions. Ainsi, le couple *directeur/entreprise* a été rapproché sur la base de contextes comme *détenir*_SUJ, *conseiller*_OBJ ou *affirmer*_SUJ, qui sont clairement destinés à être employés avec des animés. Il en va de même pour *joueur/équipe*, qui ont été rapprochés *via* les contextes *s'entraîner*_SUJ, *affronter*_SUJ ou *se qualifier*_SUJ. Nous avons ici aussi clairement affaire à un fonctionnement de type métonymique, dans la mesure où l'ensemble est

employé pour désigner les membres.

Les couples dont le méronyme est un bâtiment et l'holonyme un lieu – *château/canton*, *école/commune*, *immeuble/métropole* – sont également bien captés par l'ADA (c'est le cas 82,6 % d'entre eux). Cela peut s'expliquer par l'ambiguïté des noms de bâtiments, qui peuvent aussi bien être employés comme des noms de lieux. Ainsi, le recouvrement entre ces deux classes implique une certaine similarité au niveau des distributions de leurs membres.

À l'autre extrémité du spectre, on remarque que la catégorie LIVING_PART apparaît en position méronyme dans trois des configurations hétérogènes les moins bien repérées par l'ADA. Cette classe est successivement associée à PLANT_LIFE (*pétale/marguerite*, *tige/rose*, *tronc/chêne*), SOUL (*bras/citoyen*, *doigt/bébé*, *main/professeur*) et FAUNA (*bec/canard*, *patte/chat*, *queue/loup*). Dans les trois cas, le fait que les couples relevant de ces classes ne soient que peu repérés s'explique par le fait qu'ici, les *touts* sont des êtres animés, contrairement à leurs parties. La conséquence en est que leurs propriétés distributionnelles sont radicalement opposées à celles de leurs méronymes. Cela semble être un peu moins flagrant pour les végétaux (ce qui explique que les couples LIVING_PART/PLANT_LIFE sont mieux repérés que les couples LIVING_PART/SOUL et LIVING_PART/FAUNA). On est donc dans le cas attendu de mots relevant de sens différents et par conséquent dissemblables sur le plan distributionnel.

8.4.3 Conclusion

Dans ce chapitre, nous avons abordé la question du repérage des couples de méronymes par ADA à l'aide d'une méthode d'évaluation qualitative reposant sur l'annotation sémantique de couples de méronymes. Sur le plan méthodologique, cette étude a montré que la typologie habituellement utilisée pour décrire les différents types de relations méronymiques était peu adaptée pour catégoriser nos données. Une approche consistant à typer sémantiquement les couples de méronymes permet de mieux rendre compte de la diversité des relations qu'ils expriment. Sur le plan des résultats, nous avons montré que si la méronymie, considérée globalement, est repérée dans des proportions comparables à d'autres relations (environ un tiers des méronymes de JDM sont détectés par le programme d'ADA que nous avons utilisé), elle n'est pas repérée par l'ADA de manière homogène : la nature sémantique des mots qui entrent dans la relation de méronymie constitue un facteur décisif pour leur détection par l'ADA. Tout d'abord, nous avons constaté que l'ADA privilégiait le repérage des couples de méronymes dont les membres relèvent de la même classe sémantique. Ensuite, nous avons vu que certaines configurations étaient identifiées dans des proportions beaucoup

plus fortes que d'autres. C'est le cas des paires associant deux lieux, deux événements ou deux structures, ou associant un humain à un groupe social ou un lieu à un bâtiment. D'autres relations méronymiques, comme celles impliquant les parties du corps, sont mal détectées par l'ADA car elles mettent en jeu des termes qui ne fonctionnent pas de la même manière sur le plan distributionnel. Cette étude contribue donc à préciser les conditions d'application du critère distributionnel au repérage d'une relation sémantique donnée. À ce stade, elle laisse cependant ouverte la question de l'influence du corpus de test sur la prédominance de certaines configurations distributionnelles des résultats.

Conclusion

Quiconque est amené à consulter une ressource distributionnelle pourra témoigner du sentiment de perplexité que cela produit. D'un côté, il suffit de se rendre sur l'interface des VDW ou des VDLM et de chercher les voisins d'un mot un tant soit peu fréquent (assez pour apparaître dans la base) pour constater que les premiers voisins proposés présentent un intérêt manifeste. On peut en effet facilement identifier parmi ces voisins un vaste spectre de relations sémantiques, aussi bien classiques que non classiques. D'un autre côté, on peut se trouver rebuté par la masse de résultats fournis. En effet, si le score de similarité permet de faire apparaître en haut de liste les voisins les plus *pertinents*, ces derniers se retrouvent rapidement noyés dans une cohorte de plusieurs centaines de voisins parmi lesquels on serait bien en peine de distinguer des relations sémantiques de quelque nature que ce soit. Au vu de ce caractère pléthorique des bases distributionnelles, on comprend aisément pourquoi les méthodes d'évaluation qui sont traditionnellement appliquées à ces ressources s'en tiennent à des considérations quantitatives. Nous avons précédemment pointé du doigt les limites de ces méthodes, qui n'apportent que peu d'information sur la façon dont s'opèrent les rapprochements au sein de la base distributionnelle, préservant ainsi l'aspect *boîte noire* de l'ADA.

Tout au long de ce travail de thèse, nous avons montré l'intérêt qu'il y a à prendre le relais des approches quantitatives pour observer les voisins à l'échelle du contexte, avec le regard du linguiste. Les quatre études que nous avons menées dans les chapitres 5 à 8 ont ainsi été guidées par la question générale de savoir pourquoi certains couples de mots entretenant une relation sémantique classique ont été captés par l'ADA alors que d'autres ne l'ont pas été. Pour y répondre, nous avons mis en place quatre protocoles différents selon la nature de la relation et le phénomène que l'on a cherché à observer :

– dans le chapitre 5, nous avons montré quels étaient les phénomènes qui pouvaient être à l'origine d'un décalage distributionnel entre deux synonymes recensés dans le DES. Nous avons illustré le fait que ce

décalage s'observait de façon différente en fonction de la base de voisins utilisée. Autrement dit, le fait de filtrer le DES avec différentes bases de voisins permet d'adapter le dictionnaire à certains types de textes. Nous avons fait l'hypothèse qu'il était possible d'exploiter ce principe pour améliorer l'efficacité de la consultation du dictionnaire en proposant en priorité à l'utilisateur les synonymes qui sont les plus pertinents pour lui. Nous donnons des pistes pour un protocole expérimental qui nous permettrait de vérifier cette hypothèse ;

– dans le chapitre 6, nous avons abordé la relation d'antonymie d'un point de vue inspiré des approches psycholinguistiques. En nous appuyant sur la littérature, nous avons émis l'hypothèse selon laquelle le fait de croiser des couples de mots extraits par des patrons antonymiques et une base distributionnelle peut mettre au jour des couples antonymiques dont le double fonctionnement à la fois syntagmatique et paradigmatique serait gage d'un degré particulièrement élevé d'antonymie. Les résultats obtenus invalident cette hypothèse en montrant que le critère syntagmatique permet de faire émerger des couples d'antonymes qui ne fonctionnent pas, dans nos données, sur le mode de la substituabilité ;
– dans le chapitre 7, nous montrons l'inégalité du repérage des liens entre hyperonymes et hyponymes en fonction des bases de voisins. Nous avons distingué les cas où les hyponymes d'un mot étaient bien captés dans les trois bases, étaient captés de façon fluctuante ou étaient mal captés. Ce dernier cas de figure nous a permis de décrire plusieurs phénomènes entraînant des décalages distributionnels entre les hyponymes et les hyperonymes ;
– le chapitre 8 a consisté à mobiliser les VDW pour mettre à l'épreuve le caractère substituable des couples de méronymes. La propriété de substituabilité ne va en effet pas de soi pour un certain nombre de couples de méronymes comme *métal/épée* ou *vache/troupeau*, qui sont composés de mots appartenant à des classes sémantiques différentes. Nous avons dans un premier temps tenté d'annoter manuellement une série de couples de méronymes issus de JDM en s'appuyant sur une typologie. Les difficultés que nous avons rencontrées lors de cette tâche nous ont poussé à adopter une autre démarche consistant à assigner semi-automatiquement une classe sémantique aux mots de ces couples afin de voir si certaines combinaisons de classes étaient mieux captées par l'ADA que d'autres. Les résultats ont montré que des phénomènes comme la métonymie favorisaient le repérage de couples comme *joueur/équipe* ou *musicien/orchestre*, qui sont composés de deux mots renvoyant respectivement à un animé et à un groupe social. En revanche, les couples de méronymes comme ceux qui relient une partie

du corps et un animal – *bec/canard*, *queue/loup* – ont beaucoup moins de chances d'être captés comme des voisins. On voit donc que la nature sémantique des méronymes influe fortement sur leurs chances d'êtres captés par l'ADA. Ces résultats mettent en lumière le fait que l'ADA se montre plus efficace pour extraire certains types de méronymes que d'autres.

Le choix de nous focaliser sur l'étude de ces quatre relations a été guidé par deux critères. D'une part, nous avons voulu montrer que les apports de l'ADA pouvaient bénéficier à la description de relations déjà bien connues de la lexicologie comme la synonymie, l'antonymie, l'hyperonymie et la méronymie. D'autre part, ces relations sont parmi les seules pour lesquelles des lexiques – même imparfaits – sont disponibles. Nous avons vu au chapitre 4 que JDM contenait également de nombreux couples portant la relation IDÉE ASSOCIÉE. Ces données pourraient être utilisées dans des approches similaires à la nôtre, mais il est certain que la nature hétérogène des couples d'idées associées compliquera de façon considérable l'analyse des résultats fournis.

La question de la nature des couples contenus dans les lexiques que nous avons utilisés se pose également dans notre cas. En choisissant d'accéder aux couples de synonymes, d'hyperonymes et de méronymes présents dans les voisins *via* des lexiques, nous avons choisi de nous en remettre soit au jugement de lexicographes – pour les synonymes du DES –, soit à celui de locuteurs *tout-venant* – pour JDM – pour la sélection de nos échantillons d'étude. Nous avons par exemple eu l'occasion de commenter le caractère artificiel de certains couples de méronymes, que nous avons attribué aux mécanismes de jeu de JeuxDeMots (sections 7.2.1 et 8.3.1). Nous avons toutefois fait le choix d'utiliser cette ressource en prenant le principe de validation des couples – le fait qu'ils soient fournis par deux joueurs différents – comme un gage de leur *fiabilité*. Nous rappelons que ce choix, aussi discutable soit-il, a également été orienté par le fait que JDM était la seule ressource nous permettant de disposer de lexiques d'hyperonymes et de méronymes. Le fait de ne pas avoir approfondi la question de la qualité des lexiques utilisés peut être considéré comme une des limites de notre travail. Nous avons toutefois considéré que ce problème aurait nécessité une étude à part entière, ce qui nous aurait dévié de notre objectif initial.

Une deuxième limite liée à l'utilisation de lexiques pour étudier les voisins est que cet angle d'approche ne nous offre qu'un point de vue particulièrement réducteur sur l'étendue des relations contenues dans les voisins. Nous avons en effet montré à la section 4.1.3 que la proximité distributionnelle pouvait traduire un ensemble de relations sémantiques particulièrement hé-

térogènes (action/instrument, producteur/produit, etc.). Cette hétérogénéité complique considérablement la mise au jour du potentiel des voisins à extraire ces relations non classiques : tant que leur nature restera indéfinie, la question des méthodes qui seraient les plus adaptées pour accéder à ces relations dans les voisins restera en suspens (on pourrait voir un cercle vicieux dans le fait que la définition de ces relations nécessiterait dans un premier temps de pouvoir en observer des occurrences dans les voisins, donc de trouver un moyen de les extraire...).

Comme nous l'avons indiqué dans l'introduction de ce document, en abordant ces objets que sont les bases distributionnelles avec une perspective linguistique, nous nous inscrivons dans une approche dont les apports peuvent bénéficier à deux types de publics :

- pour les spécialistes du TAL, familiers des méthodes distributionnelles, nos travaux apportent un nouveau regard sur une problématique déjà bien connue. En effet, plutôt que de s'en tenir à une comparaison des ressources générées automatiquement avec des lexiques, notre approche a consisté à comparer les contextes d'apparition des couples de ces lexiques dans les corpus. Nous avons ainsi montré l'influence de certains facteurs linguistiques sur la distribution des mots et donc sur la nature des couples extraits par l'ADA. Ces résultats plaident ainsi pour une évaluation *raisonnée* des ressources distributionnelles inspirée des méthodes de la linguistique de corpus qui pourrait être menée en complément d'une approche quantitative ;
- pour les lexicologues n'ayant jamais eu l'occasion de manipuler une base distributionnelle, nos résultats se veulent une démonstration du potentiel de ces ressources pour la description des relations lexicales. En effet, les relations lexicales se définissent traditionnellement en fonction du critère de substituabilité. Le fait de disposer d'une ressource qui consiste en la mise en œuvre à grande échelle de ce principe de substituabilté offre ainsi la possibilité de mettre à l'épreuve un certain nombre de principes considérés comme acquis. Nous avons par exemple montré qu'un certain nombre de couples de synonymes, d'hyperonymes ou de méronymes contenus dans les lexiques, donc recueillis *in abstracto*, ne fonctionnaient pas sur le mode de la substituabilité dans nos corpus (soit dans les trois que nous avons étudiés, soit seulement dans certains d'entre eux). En définitive, au delà de l'intérêt des phénomènes qu'elles ont permis de mettre au jour, ces études ont surtout cherché à convaincre les linguistes de s'approprier ces bases distributionnelles, qui constituent à la fois un outil permettant de vérifier des hypothèses

formulées par ailleurs et un objet d'étude à part entière.

Les perspectives que nous envisageons pour la suite de nos travaux de recherche se situent à deux niveaux. D'une part, nous avons pour projet de mettre en place un protocole d'évaluation afin de tester l'hypothèse que nous avons développée dans le chapitre 5. Nous avons supposé que le fait de filtrer les synonymes du DES avec des bases distributionnelles permettrait de réorganiser le dictionnaire de façon à présenter à l'utilisateur les synonymes qui sont les plus pertinents pour lui. Le protocole consistera en un questionnaire dans lequel il sera demandé à une série de participants de classer ou de noter les différents synonymes d'un mot cible en fonction du contexte dans lequel il se trouve. Ce contexte sera une phrase entière, extraite du corpus qui a permis de générer la base de voisin avec laquelle ont été croisés les synonymes. Si les utilisateurs considèrent comme plus pertinents les synonymes du mot cible qui ont été captés par l'AD, alors notre hypothèse sera vérifiée. Comme nous l'avons indiqué, à la fin du chapitre 5, il reste de nombreuses questions à éclaircir sur le protocole (formulation de la consigne, présentation des synonymes, etc.).

Un autre prolongement de ces travaux consisterait à appliquer notre mode d'évaluation aux bases distributionnelles qui ont été récemment générées au sein du laboratoire CLLE-ERSS à l'aide de l'analyseur syntaxique Talismane (cf. section 3.2.2). Dans le travail que nous avons mené ici, les trois ressources distributionnelles que nous avons eu l'occasion de comparer ont toutes été générées par la chaîne Syntex-Upéry. Dès lors, le fait de pouvoir mettre face à face deux ressources générées à partir de corpus identiques mais avec des dispositifs distincts nous permettrait de pousser la démarche comparative dans une direction différente. Cela nous donnerait en effet la possibilité de mener une évaluation réciproque des ressources qui pourrait nous permettre d'analyser l'influence des chaînes de traitement sur la nature des rapprochements générés.

Dans cette optique, nous nous positionnons en faveur d'une démarche qui consisterait à générer une variété de ressources selon des paramètres différents. Le fait de faire varier au maximum la nature des corpus et les paramètres des chaînes de traitement offrirait en effet un terrain particulièrement riche pour la démarche d'évaluation : il serait alors possible d'isoler certains facteurs et d'observer leur influence sur la nature des couples mis au jour. L'objectif de cette démarche serait d'avoir une meilleure connaissance des *leviers* qui sont manipulés lors de la génération de ressources distributionnelles pour, à terme, avoir assez de contrôle sur la chaîne de traitement pour pouvoir générer les ressources qui répondront le mieux à des besoins formulés *a priori*.

Index

Table des figures

Liste des tableaux

Liste des formules

Bibliographie

Abbès, S. B., Zargayouna, H. et Nazarenko, A. (2011). Evaluating semantic classes used for ontology building and learning from texts. *In KEOD*, pages 445–448.

Adam, C. (2012). *Voisinage lexical pour l'analyse du discours*. Thèse de doctorat. Université Toulouse II – Le Mirail.

Adam, C. et Morlane-Hondère, F. (2009). Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique. *In Actes de la conférence RECITAL'09*, page article 6, Senlis, France.

Adda, G., Sagot, B., Fort, K. et Mariani, J. (2011). Crowdsourcing for Language Resource Development : Critical Analysis of Amazon Mechanical Turk Overpowering Use. *In LTC 2011 : Proceedings of the 5th Language and Technology Conference*, Poznan, Pologne.

Alario, F.-X. et Ferrand, L. (1998). Normes d'associations verbales pour 366 noms d'objets concrets. *L'année psychologique*, 98(4):659–709.

Aussenac-Gilles, N., Biébow, B. et Szulman, S. (2003). D'une méthode à un guide pratique de modélisation de connaissances à partir de textes. *In 5e rencontres Terminologie et Intelligence Artificielle (TIA 2003)*, pages 41–53, Strasbourg, France. ENSSAIS.

Baker, K. (2005). Singular value decomposition tutorial, `http://www.ling.ohio-state.edu/~kbaker/`.

Bally, C. (1951). *Traité de stylistique française*. Numéro 1 de Traité de stylistique française. Librairie de l'Université, Genève.

Banerjee, S. et Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810.

Bannour, S., Audibert, L. et Nazarenko, A. (2011). Mesures de similarité distributionnelle entre termes. *In 22es journées francophones d'ingénierie des connaissances*, pages 523–538, Chambéry, France.

Baroni, M., Bernardini, S., Ferraresi, A. et Zanchetta, E. (2009). The WaCky wide web : a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M. et Lenci, A. (2010). Distributional Memory : A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, M. et Lenci, A. (2011). How we blessed distributional semantic evaluation. *In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS'11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baroni, M. et Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices : representing adjective-noun constructions in semantic space. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barsalou, L. W. (1983). Ad hoc categories. *Memory and cognition*, 11(3): 211–227.

Bloomfield, L. (1935). *Language*. Allen and Unwin, London.

Bloomfield, L. (1970). *Le langage*. Bibliothèque scientifique. Payot, Paris.

Bourigault, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *TAL. Traitement automatique des langues*, 34(2):105–117.

Bourigault, D. (1994). *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*. Thèse de doctorat, Ecole des Hautes Etudes en Sciences Sociales, Paris.

Bourigault, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle* (24–27 juin 2002), Nancy.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel :* Syntex. Habilitation à diriger des recherches. Université Toulouse II – Le Mirail.

Bourigault, D., Aussenac-Gilles, N. et Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18(1):87–110.

Bourigault, D. et Galy, E. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *In 4e Journées de la linguistique de corpus* (15–17 septembre 2005), Lorient.

Bourigault, D. et Lame, G. (2002). Analyse distributionnelle et structuration de terminologie. Application a la construction d'une ontologie documentaire du Droit. *Traitement automatique des langues*, 43(1).

Brousseau, A. et Roberge, Y. (2000). *Syntaxe et sémantique du français*. Champs Linguistiques Series. Fides.

Budanitsky, A. et Hirst, G. (2001). Semantic distance in WordNet : an experimental, application-oriented evaluation of five measures. *In Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, USA.

Budanitsky, A. et Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Bullinaria, J. et Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics : a computational study. *Behavior Research Methods*, 39(3).

Bullinaria, J. et Levy, J. (2012). Extracting semantic representations from word co-occurrence statistics : stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3).

Buvet, P.-A. et Grezka, A. (2009). Les dictionnaires électroniques du modèle des classes d'objets. *In* Blanco, X. et Buvet, P.-A., éditeurs : *Les représentations des structures prédicat-arguments*, pages 63–79.

Chaffin, R. et Herrmann, D. J. (1984). The similarity and diversity of semantic relations. *Memory and Cognition*, 12(2):134–141.

Charles, W. et Miller, G. (1989). Context of antonymous adjectives. *Applied psycholinguistics*, 10.

Church, K. W. et Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Clarke, D. (2012). A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.

Croft, W. et Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Cruse, D. A. (2002). Hyponymy and its varieties. *In* Green, R., Bean, C. et Myaeng, S., éditeurs : *The Semantics of Relationships*, pages 35–50.

Cruse, D. A. (2004). *Meaning in Language : An Introduction to Semantics and Pragmatics*. Oxford Textbooks in Linguistics Series. Oxford University Press.

Cruse, D. A. (2006). *A Glossary of Semantics and Pragmatics*. Glossaries in Linguistics Series. Columbia University Press.

Curran, J. R. (2004). *From distributional to semantic similarity*. Thèse de doctorat, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh.

Curran, J. R. et Moens, M. (2002). Scaling context space. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 231–238, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dagan, I., Lee, L. et Pereira, F. (1997). Similarity-based methods for word sense disambiguation. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 56–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dagan, I., Marcus, S. et Markovitch, S. (1993). Contextual word similarity and estimation from sparse data. *In* Schubert, L. K., éditeur : *ACL*, pages 164–171.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *LLC*, 25(4):447–464.

de Saussure, F. (1916). *Cours de linguistique générale*. Bayot, Paris.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. et Harshman, R. (1990). Indexing by latent semantic analysis. *Journal Of The American Society For Information Science*, 41(6):391–407.

Deese, J. (1964). The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3(5):347–357.

Deese, J. (1965). *The structure of associations in language and thought*. Johns Hopkins Press.

Dubois, J. (1969). Grammaire distributionnelle. *Langue française*, 1(1):41–48.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Duscherer, K. et Mounoud, P. (2006). Normes d'associations verbales pour 151 verbes d'action. *L'année psychologique*, 106:397–413.

Edmonds, P. (1999). *Semantic Representations of Near-Synonyms for Automatic Lexical Choice*. Thèse de doctorat.

Erk, K. et Padó, S. (2008). A structured vector space model for word meaning in context. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 897–906, Stroudsburg, PA, USA. Association for Computational Linguistics.

Evert, S. (2004). *The Statistics of Word Cooccurrences : Word Pairs and Collocations*. Thèse de doctorat, Université de Stuttgart.

Evert, S. (2008). Corpora and collocations. *In* Lüdeling, A. et Kytö, M., éditeurs : *Corpus Linguistics. An International Handbook*.

Evert, S., Baroni, M. et Lenci, A. (2010). *Distributional Semantic Models*. Tutorial at NAACL-HLT 2010, Los Angeles, CA.

Evert, S. et Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. *In In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195.

Evert, S. et Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, 19(4):450–466.

Fabre, C. (2010). *Affinités syntaxiques et sémantiques entre mots : apports mutuels de la linguistique et du TAL*. Hdr, Université Toulouse le Mirail - Toulouse II.

Fabre, C. et Habert, B. (1998). Acquisition de relations entre mots pour une lecture sémantique de corpus. *In* Mellet, S., éditeur : *4èmes journées internationales d'analyse statistique des données textuelles (JADT'98)*, pages 273–282, Nice.

Fabre, C., Habert, B. et Labbé, D. (1997). La polysémie dans la langue générale et les discours spécialisés. *Sémiotiques*, (13):15–31.

Faure, D. et Nédellec, C. (1998). ASIUM : learning subcategorization frames and restrictions of selection. *In* Kodratoff, Y., éditeur : *10th European Conference on Machine Learning (ECML 98) – Workshop on Text Mining*, Chemnitz Allemagne.

Fellbaum, C. (1995). Co-occurrence and antonymy. *International Journal of Lexicography*, 8.

Fellbaum, C., éditeur (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Ferrand, L. (2001). Normes d'associations verbales pour 260 mots "abstraits". *L'année psychologique*, 101(4):683–721.

Ferrara, A. (2010). Les dictionnaires de synonymes : une typologie évoluant avec le temps. *In* Neveu, F., Muni Toke, V., Klingler, T., Durand, J., Mondada, L. et Prévost, S., éditeurs : *2ème Congrès Mondial de Linguistique Française*, page 62.

Ferret, O. (2010). Similarité sémantique et extraction de synonymes à partir de corpus. *In Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Canada. ATALA.

Finkelstein, L., Evgenly, G., Yossi, M., Ehud, R., Zach, S., Gadi, W. et Eytan, R. (2001). Placing search in context : the concept revisited. *In Proceedings of the Tenth International World Wide Web Conference*.

Firth, J. R. (1957). *Papers in linguistics 1934-1951*. Oxford University Press, London.

Fišer, D. et Sagot, B. (2008). Combining multiple resources to build reliable wordnets. *In TSD*, pages 61–68.

Fletcher, W. H. (2011). Corpus analysis of the World Wide Web. *In* Chapelle, C. A., éditeur : *Encyclopedia of applied linguistics*. Wiley-Blackwell, Oxford.

Foltz, P., Kintsch, W. et Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3):285–307.

Francis, W. N. et Kucera, H. (1979). Brown Corpus Manual. Rapport technique, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Frege, G. (1884). *Grundlagen der Arithmetik : Eine logisch mathematische Untersuchung über den Begriff der Zahl*. Wilhelm Koebner, Breslau.

Fung, P. et McKeown, K. (1997). Finding terminology translations from non-parallel corpora. *In Proc. of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Gale, W. A., Church, K. W. et Yarowsky, D. (1994). Discrimination Decisions for 100,000 – Dimensional Spaces. *In Journal of Operations Research*, pages 429–450. Kluwer Academic Publishers.

Gamallo Otero, P. (2008). Comparing Window and Syntax Based Strategies for Semantic Extraction. *In* Teixeira, A. J. S., de Lima, V. L. S., de Oliveira, L. C. et Quaresma, P., éditeurs : *PROPOR*, volume 5190 de *Lecture Notes in Computer Science*, pages 41–50. Springer.

Giesbrecht, E. et Evert, S. (2009). Part-of-speech tagging – a solved task ? An evaluation of POS taggers for the Web as corpus. *In* Alegria, I., Leturia, I. et Sharoff, S., éditeurs : *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain.

Grefenstette, G. (1992a). Finding semantic similarity in raw text : the Deese antonyms. *In Fall Symposium Series, Working Notes, Probabilistic Approaches to Natural Language*, pages 61–65.

Grefenstette, G. (1992b). Sextant : exploring unexplored contexts for semantic extraction from syntactic analysis. *In proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 324–326, Newark, Delaware, USA. Association for Computational Linguistics.

Grefenstette, G. (1994a). Corpus-Derived First, Second and Third-Order Word Affinities. *In Proceedings of EURALEX'94*.

Grefenstette, G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

Grefenstette, G. (1996). Evaluation techniques for automatic semantic extraction : comparing syntactic and window based approaches. *In* Boguraev, B. et Pustejovsky, J., éditeurs : *Corpus processing for lexical acquisition*, pages 205–216. MIT Press, Cambridge, MA, USA.

Habert, B. (1998). *Des mots complexes possibles aux mots complexes existants : l'apport des corpus*. Habilitation à diriger des recherches en linguistique, Université Lille III – Charles de Gaulle.

Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Collection l'Essentiel Français. Ophrys.

Habert, B. et Nazarenko, A. (1996). La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience. *In Journées sur l'acquisition des connaissances*, pages 137–142, Sète. AFIA.

Habert, B. et Zweigenbaum, P. (2002). Contextual Acquisition of Information Categories : what has been done and what can be done automatically ? *In* Nevin, B., éditeur : *The Legacy of Zellig Harris : Language and information into the 21st century*. John Benjamins.

Habert, B. et Zweigenbaum, P. (2003). Classer les mots : sémantique à gros grain et méthodologie harrissienne. *Revue de Sémantique et Pragmatique*, 12:101–119.

Halliday, M. A. K. et Hasan, R. (1976). *Cohesion in English*. Longman Publishing Group.

Harper, K. E. (1961). *Procedures for the Determination of Distributional Classes*. Memorandum. Rand Corporation.

Harper, K. E. (1965). Measurement of similarity between nouns. *In Proceedings of the 1965 conference on Computational linguistics*, COLING '65, pages 1–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Harris, Z. (1968). Mathematical structures of language. John Wiley & Sons.

Harris, Z. (1991). *A theory of language and information : a mathematical approach*. Clarendon Press ; Oxford University Press, Oxford, England.

Harris, Z., Gottfried, M., Ryckman, T., Mattcik, Jr., P., Daladier, A., Harris, T. N. et Harris, S. (1989). *The Form of Information in Science : Analysis of an Immunology Sublanguage*. Kluwer Academic Publishers, Dordrecht.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. *In Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Henestroza Anguiano, E. et Denis, P. (2011). FreDist : Automatic construction of distributional thesauri for French. *In Actes de la 18ème conférence sur le traitement automatique des langues naturelles*, pages 119–124, Montpellier, France, France.

Herrmann, D. J., Chaffin, R., Daniel, M. P. et Wool, R. S. (1986). The role of elements of relation definition in antonymy and synonym comprehension. *Zeitschrift fur Psychologie*, 194.

Hindle, D. (1990). Noun classification from predicate-argument structure. *In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Hirschman, L., Grishman, R. et Sager, N. (1975). Grammatically-based automatic word class formation. *Information Processing and Management*, 11(1-2):39–57.

Hoey, M. (1991). *Patterns of lexis in text*. Describing English language. Oxford University Press.

Ide, N. et Véronis, J. (1998). Introduction to the special issue on word sense disambiguation : the state of the art. *Computational Linguistics*, 24(1):2–40.

Inkpen, D. (2003). *Building a Lexical Knowledge-Base of Near-Synonym Differences*. Thèse de doctorat, Université de Toronto.

Jackson, H. (2002). *Lexicography : an introduction*. Taylor & Francis Group.

Jones, S. (2002). *Antonymy : A Corpus-Based Perspective*. Routledge Advances in Corpus Linguistics. Taylor & Francis.

Jones, S., Murphy, L., Paradis, C. et Willners, C. (2012). *Antonyms in English : Construals, Constructions and Canonicity*. Studies in English Language. Cambridge University Press.

Jones, S., Paradis, C., Murphy, M. L. et Willners, C. (2007). Googling for 'opposites' : a web-based study of antonym canonicity. *Corpora*, 2(2): 129–154.

Joos, M. (1950). Description of language design. *The Journal of the Acoustical Society of America*, (22):701–708.

Joubarne, C. et Inkpen, D. (2011). Comparison of semantic similarity for different languages using the google n-gram corpus and second- order co-occurrence measures. *In Proceedings of the 24th Canadian conference on Advances in artificial intelligence*, Canadian AI'11, pages 216–221, Berlin, Heidelberg. Springer-Verlag.

Justeson, J. et Katz, S. (1991). Co-occurrence of antonymous adjectives and their contexts. *Computational linguistics*, 17.

Keet, C. et Artale, A. (2008). Representing and reasoning over a taxonomy of part–whole relations. *Applied Ontology*, 3(1):91–110.

Kerbrat-Orecchioni, C. (1977). *La connotation*. Linguistique et sémiologie. Presses universitaires de Lyon.

Kilgarriff, A. et Yallop, C. (2000). What's in a thesaurus ? *In LREC*. European Language Resources Association.

Kittredge, R. (1982). Variation and homogeneity of sublanguages. *In* Kittredge, R. et Lehrberger, J., éditeurs : *Sublanguage : Studies of Language in Restricted Domains*, pages 107–137. de Gruyter, Berlin/New York.

Kittredge, R. (2003). Sublanguages and controlled languages. *In The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Kleiber, G. et Tamba, I. (1990). L'hyponymie revisitée : inclusion et hiérarchie. *Langages*, 25(98):7–32.

Kolb, P. (2008). DISCO : A Multilingual Database of Distributionally Similar Words. *In* Storrer, A., Geyken, A., Siebert, A. et Würzner, K.-M., éditeurs : *KONVENS 2008 – Ergänzungsband : Textressourcen und lexikalisches Wissen*, pages 37–44.

Köhler, W. (1929). *Gestalt Psychology*. Traduction française, *La psychologie de la forme*, Gallimard, Paris, 1964.

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. *In SNLP'07 : 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand.

Lakoff, G. (1987). *Women, fire, and dangerous things : what categories reveal about the mind*. Cognitive science/linguistics/philosophy. University of Chicago Press.

Landauer, T. et Dumais, S. (1997). A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

Landis, J. R. et Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Le Guern, M. (1972). *Sémantique de la métaphore et de la métonymie*. Langue et Langage. Librairie Larousse.

Le Moigno, S., Charlet, J., Bourigault, D. et Jaulent, M. C. (2002). Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale. *In Actes des 13es journées francophones d'ingénierie des connaissances (IC 2002)*, pages 229–238, Rouen. Morgan Kaufmann.

Lee, L. (1999). Measures of distributional similarity. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lehmann, A. et Martin-Berthet, F. (2011). *Introduction à la lexicologie : Sémantique et morphologie*. Collection Lettres supérieures. Armand Colin.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. *In proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA. Association for Computational Linguistics.

Lin, D. (1998b). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

Lin, D., Zhao, S., Qin, L. et Zhou, M. (2003). Identifying synonyms among distributionally similar words. *In Proceedings of the 18th international joint conference on Artificial intelligence*, IJCAI'03, pages 1492–1493, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lindsey, R., Veksler, V., Grintsvayg, A. et Gray, W. (2007). Be wary of what your computer reads : the effects of corpus selection on measuring semantic relatedness. *In 8th International Conference of Cognitive Modeling, ICCM*.

Ljubešić, N., Boras, D., Bakarić, N. et Njavro, J. (2008). Comparing measures of semantic similarity. *In Proceedings of the 30th International Conference on Information Technology Interfaces*.

Lobanova, A. (2012). *The anatomy of antonymy : a corpus-driven approach*. Thèse de doctorat, University of Groningen, The Netherlands.

Lobanova, A., van der Kleij, T. et Spenader, J. (2010). Defining Antonymy : A Corpus-based Study of Opposites by Lexico-syntactic Patterns. *International Journal of Lexicography*, 23(1):19–53.

Lund, K. et Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28:203–208.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. University Press, Cambridge.

Lyons, J. (1977). *Semantics*. Semantics. Cambridge University Press.

Lyons, J. (1995). *Linguistic Semantics : An Introduction*. Cambridge University Press.

Manguin, J.-L. (2002). Le dictionnaire électronique des synonymes du CRISCO. *In Colloques Sciences humaines et nouvelles technologies*, Tunis Tunisia.

Manguin, J.-L. (2005). Les dictionnaires en ligne : nouvelles diffusions, nouveaux objectifs. Paris, France.

Manguin, J. L., François, J., Eufe, R., Fesenmeier, L., Ozouf, C. et Sénéchal, M. (2004). Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux. *In Cahiers du CRISCO*, volume 34. CRISCO, Université de Caen.

Manning, C. D., Raghavan, P. et Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Manning, C. D. et Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Marcu, D. et Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 368–375, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martinot, C. (2007). Les langues de spécialité en question : perspectives d'étude et applications. *In Présentation de la 12e Journée Scientifique de la Cellule de Recherche en Linguistique*, Université Paris Diderot.

Mel'čuk, I. (1988). *Dependency Syntax : Theory and Practice*. State University of New York Press.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S. et Polguère, A. (1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV*. Presses Universitaires de Montréal.

Memmi, D. (2000). Le modèle vectoriel pour le traitement de documents. *Cahiers Leibniz*, (14).

Miller, G. A. et Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mitchell, J. et Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Mohammad, S. M., Dorr, B., Hirst, G. et Turney, P. D. (2013). Computing lexical contrast. À paraître in *Computational Linguistics*.

Mondary, T. (2011). *Construction d'ontologies à partir de textes. L'apport de l'analyse de concepts formels.* Thèse de doctorat, Université Paris-Nord - Paris XIII.

Morin, E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse de doctorat, Université de Nantes.

Morlane-Hondère, F. et Fabre, C. (2012). Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. *SHS Web of Conferences*, 1:1001–1015.

Morris, J. (2007). *Readers' Perceptions of Lexical Cohesion and Lexical Semantic Relations in Text*. Thèse de doctorat, University of Toronto (Canada).

Morris, J. et Hirst, G. (2004). Non-classical lexical semantic relations. *In Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS '04, pages 46–51, Stroudsburg, PA, USA. Association for Computational Linguistics.

Murphy, L. (2006). Antonyms as lexical constructions : or, why paradigmatic construction is not an oxymoron. *Constructions all over : case studies and theoretical implications.* Special volume of *Constructions*, SV1(8).

Murphy, M. L. (2003). *Semantic Relations and the Lexicon : Antonymy, Synonymy and other Paradigms*. University Press, Cambridge.

Nazarenko, A. (2004). Donner accès au contenu des documents textuels : acquisition de connaissances et analyse de corpus spécialisés. Mémoire d'Habilitation à Diriger des Recherches en Informatique. Université Paris 13.

Nazarenko, N., Zweigenbaum, P., Habert, B. et Bouaud, J. (2001). Corpus-based extension of a terminological semantic lexicon. *In* Bourigault, D., Jacquemin, C. et L'Homme, M.-C., éditeurs : *Recent Advances in Computational Terminology*, Natural Language Processing, chapitre 16, pages 327–351. John Benjamins, Amsterdam.

Orosemane, L. (2012). Qui se cache derrière le dictionnaire des synonymes de Caen ? *Rue 89*. En ligne. `www.rue89.com/2012/10/28/qui-se-cache-derriere-le-dictionnaire-des-synonymes-de-caen-236552` Page consultée le 25/02/2013.

Padó, S. et Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

Panchenko, A. et Morozova, O. (2012). A study of hybrid similarity measures for semantic relation extraction. *In Proceedings of the Workshop*

on Innovative Hybrid Approaches to the Processing of Textual Data, pages 10–18, Avignon, France. Association for Computational Linguistics.

Pantel, P. et Lin, D. (2002). Discovering word senses from text. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 613–619, New York, NY, USA. ACM.

Paroubek, P., Robba, I., Vilnat, A. et Ayache, C. (2006). Data, Annotations and Measures in EASY – the Evaluation Campaign for Parsers of French. *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 315–320.

Patel, M., Bullinaria, J. A. et Levy, J. P. (1997). Extracting semantic representations from large text corpora. *In Proceedings of the 4th Neural Computation and Psychology Workshop*, pages 199–212. Springer.

Peirsman, Y., De Deyne, S., Heylen, K. et Geeraerts, D. (2008). The construction and evaluation of word space models. *In* Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S. et Tapias, D., éditeurs : *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Peirsman, Y., Heylen, K. et Speelman, D. (2007). Finding semantically related words in Dutch. co-occurrences versus syntactic contexts. *In CoSMO Workshop*, pages 9–16.

Pekar, V. (2004). Linguistic preprocessing for distributional classification of words. *In* Zock, M., éditeur : *COLING 2004 Enhancing and using electronic dictionaries*, pages 15–21, Geneva, Switzerland. COLING.

Pereira, F. C. N., Tishby, N. et Lee, L. (1993). Distributional Clustering of English Words. *In Meeting of the Association for Computational Linguistics*, pages 183–190.

Petit, G. (2005). Synonymie et dénomination. *In Linx.* En ligne. 52. Mis en ligne le 27 janvier 2011, consulté le 11 mars 2013. `http://linx.revues.org/198`.

Poibeau, T., Dutoit, D. et Bizouard, S. (2002). Évaluer l'acquisition semi-automatique de classes sémantiques. *In Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle* (24-27 juin 2002), Nancy.

Poibeau, T. et Messiant, C. (2008). Do we still Need Gold Standards for Evaluation ? *In Proceedings of the Language Resource and Evaluation Conference*, Maroc.

Polguère, A. (1998). La Théorie Sens-Texte. *In Dialangue*, volume 8-9, pages 9–30. Université du Québec à Chicoutimi.

R Development Core Team (2011). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rapp, R. (2002). The computation of word associations : comparing syntagmatic and paradigmatic approaches. *In Proceedings of the 19th international conference on Computational linguistics*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Resnik, P. (1993). *Selection and Information : A Class-Based Approach to Lexical Relationships*. Thèse de doctorat, The Institute For Research In Cognitive Science, University of Pennsylvania.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *In IJCAI*, pages 448–453. Morgan Kaufmann.

Riloff, E. et Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124.

Roark, B. et Charniak, E. (1998). Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 1110–1116, Montreal, Canada.

Rothenhäusler, K. et Schütze, H. (2009). Unsupervised classification with dependency based word spaces. *In Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rubenstein, H. et Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.

Ruge, G. (1992). Experiments on linguistically-based term associations. *Information Processing and Management*, 28(3):317–332.

Ruge, G. (1995). Human memory models and term association. *In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 219–227, New York, NY, USA. ACM.

Sager, N. (1986). Sublanguage : Linguistic phenomenon, computational tool. *In* Kittredge, R. G. . R., éditeur : *Analyzing Language in Restricted Domains : Sublanguage Description and Processing*, pages 1–18. Lawrence Erlbaum Associates.

Sager, N. et Ngô Thanh, N. (2002). The computability of strings, transformations, and sublanguage. *In* Nevin, B., éditeur : *The Legacy of Zellig Harris : Language and information into the 21st Century*, pages 79–120. John Benjamins.

Sahlgren, M. (2006). *The Word-Space Model : using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Thèse de doctorat, Stockholm University.

Sahlgren, M. (2008). The distributional hypothesis. *Rivista di Linguistica*, 20(1):33–53.

Sahlgren, M. et Karlgren, J. (2005). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3):327–341.

Sajous, F. (2009). *Documentation Upery.* Document interne au laboratoire CLLE-ERSS (Université de Toulouse).

Salton, G. et McGill, M. J. (1983). *Introduction to modern information retrieval.* McGraw-Hill computer science series. McGraw-Hill.

Salton, G., Wong, A. et Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of International Conference on New Methods in Language Processing.*

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Sparck Jones, K. et Galliers, J. R., éditeurs (1996). *Evaluating Natural Language Processing Systems : An Analysis and Review*, volume 1083 de *Lecture Notes in Computer Science*. Springer.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Storjohann, P. (2005). Corpus-driven vs. corpus-based approach to the study of relational patterns. *In Proceedings of the Corpus Linguistics conference*, Birmingham.

Terra, E. et Clarke, C. L. A. (2003). Frequency estimates for statistical word similarity measures. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 165–172, Stroudsburg, PA, USA. Association for Computational Linguistics.

Theissen, A. (1997). *Le choix du nom en discours*. Langue & cultures. Librairie Droz.

Turney, P. D. (2001). Mining the Web for Synonyms : PMI-IR versus LSA on TOEFL. *In Proceedings of the 12th European Conference on Machine Learning*, pages 491–502. Springer-Verlag.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. *In COLING*, pages 905–912.

Turney, P. D. et Pantel, P. (2010). From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research (JAIR)*, 37:141–188.

Tutin, A. (2007). Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. *In Actes de TALN 2007*, pages 283–292. Communication affichée.

Urieli, A. et Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane (à paraître). *In Actes de TALN 2013*.

Van Campenhoudt, M. (1996). Recherche d'équivalences et structuration des réseaux notionnels : le cas des relations méronymiques. *Terminology*, 3(1):53–83.

van de Cruys, T. (2010). *Mining for Meaning. The Extraction of Lexico-Semantic Knowledge from Text*. Thèse de doctorat, University of Groningen, The Netherlands.

van der Plas, L. et Bouma, G. (2004). Syntactic contexts for finding semantically related words. *In* van der Wouden, T., Poss, M., Reckman, H. et Cremers, C., éditeurs : *CLIN*. LOT Utrecht.

van der Plas, L. E. (2008). *Automatic Lexico-semantic Acquisition for Question Answering*. Thèse de doctorat. Université de Groningen.

van der Plas, L. E. (2009). Combining syntactic co-occurrences and nearest neighbours in distributional methods to remedy data sparseness. *In Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, UMSLLS '09, pages 45–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vergez-Couret, M. et Adam, C. (2012). Signaling Elaboration : Combining French Gerund Clauses with Lexical Cohesion Cues. *Discours*, 10:http ://discours.revues.org/8631.

Vossen, P., éditeur (1998). *EuroWordNet : a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Véronis, J. (2003). Cartographie lexicale pour la recherche d'information. *In* Daille, B., éditeur : *Actes de TALN 2003*, pages 97–123.

Wandmacher, T., Ovchinnikova, E. et Alexandrov, T. (2008). Does latent semantic analysis reflect human associations ? *In Proceedings of the Lexical Semantics workshop at ESSLLI'08*, Hambourg.

Weeds, J. (2003). *Measures and applications of lexical distributional similarity*. Thèse de doctorat, Université du Sussex.

Weeds, J. et Weir, D. (2005). Co-occurrence retrieval : A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.

Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. *In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 276–283, Edmonton, Canada.

Widdows, D. (2008). Semantic vector products : Some initial investigations. *In Proceedings of the Second AAAI Symposium on Quantum Interaction*.

Wiemer-Hastings, P. et Zipitria, I. (2001). Rules for syntax, vectors for semantics. *In Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*, pages 1112–1117. Erlbaum.

Willners, C. et Paradis, C. (2010). Swedish opposites : A multi-method approach to goodness of antonymy. *In Lexical-Semantic Relations : Theoretical and Practical Perspectives*. John Benjamins Publishing Company.

Winston, M. E., Chaffin, R. et Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell, Oxford. Traduit par G.E.M. Anscombe.

Zampa, V. et Lafourcade, M. (2011). PtiClic et PtiClic-Kids : jeux avec les mots permettant une acquisition lexicale par le joueur et par la machine. *Revue STICEF*, 18.

Zargayouna, H. et Nazarenko, A. (2010). Evaluation of textual knowledge acquisition tools : a challenging task. *In LREC*.

Zweigenbaum, P. et Consortium MENELAS (1994). MENELAS : accès à des comptes-rendus d'hospitalisation en langage naturel. *In* Scherrer, J.-R., éditeur : *5èmes Journées Francophones d'Informatique Médicale*, pages 23–30, Genève.

Zweigenbaum, P. et Habert, B. (2006). Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. *Glottopol*, 8:22–44.